WADD TECHNICAL REPORT 60-661

AD 249 268

TR-01-354

DISTRIBUTION-FREE STATISTICAL TESTS

James V. Bradley

Behavioral Sciences Laboratory
Aerospace Medical Division

AUGUST 1960

WRIGHT AIR DEVELOPMENT DIVISION

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

☆

Qualified requesters may obtain copies of this report from the Armed Services Technical Information Agency, (ASTIA), Arlington Hall Station, Arlington 12, Virginia.

☆

This report has been released to the Office of Technical Services, U. S. Department of Commerce, Washington 25, D. C., for sale to the general public.

☆

Copies of WADD Technical Reports and Technical Notes should not be returned to the Wright Air Development Division unless return is required by security considerations, contractual obligations, or notice on a specific document.

AD-249268

WADD TECHNICAL REPORT 60-661

DISTRIBUTION-FREE STATISTICAL TESTS

James V. Bradley

Behavioral Sciences Laboratory
Aerospace Medical Division

AUGUST 1960

Project No. 7184
Task No. 71581

WRIGHT AIR DEVELOPMENT DIVISION
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

1,000 — November 1960 — 10-465

FOREWORD

This report was prepared by the Engineering Psychology Branch of the Behavioral Sciences Laboratory, Aerospace Medical Division, Wright Air Development Division. The work began as the responsibility of the Controls Section, under Research and Development Task Number 7182 - 71514 with James V. Bradley acting as Task Scientist. It continued under subsequent Research and Development Task 7184 - 71581 and was finished by the author as a member of the Maintenance Design Section. The manuscript was typed at the Aviation Psychology Project, Miami University, under Contract Number AF 33(616)-5624, under the technical supervision of Dr. Clarke W. Crannell and Dr. S. A. Switzer.

The material included is the result of a review of the literature begun early in 1955 with the approval of Mr. John W. Senders, then Section Chief of the Controls Section, and ending early in 1958. The author was greatly aided in this effort by I. R. Savage's "Bibliography of Nonparametric Statistics and Related Topics", by hundreds of statisticians and institutions sending reprints and by the encouragement of his colleagues. He is particularly indebted to Dr. Philburn Ratoosh who critically reviewed the next-to-final draft, to Dr. Virginia L. Senders and Dr. Harry J. Jerison whose constant interest helped the author to maintain momentum, and to Mr. John W. Senders, Dr. H. R. van Saun, Dr. John P. Hornseth and Major Leroy Pigg who, as Section Chiefs, exercised their administrative powers in support of the undertaking.

ABSTRACT

As a result of an extensive survey of the literature, a large number of distribution-free statistical tests are examined. Tests are grouped together primarily according to general type of mathematical derivation or type of statistical "information" used in conducting the test. Each of the more important tests is treated under the headings: Rationale, Null Hypothesis, Assumptions, Treatment of Ties, Efficiency, Application, Discussion, Tables, and Sources. Derivations are given and mathematical interrelationships among the tests are indicated. Strengths and weaknesses of individual tests, and of distribution-free tests as a class compared to parametric tests, are discussed.

PUBLICATION REVIEW

Walter F. Grether

WALTER F. GRETHER
Technical Director
Behavioral Sciences Laboratory
Aerospace Medical Division

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

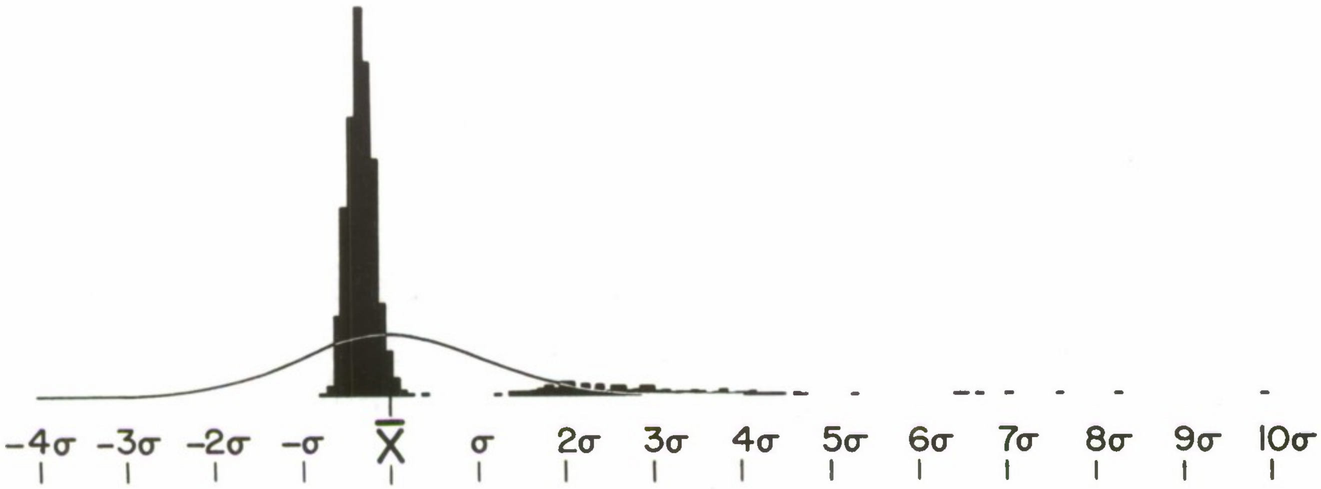


Figure 1. Radically nonnormal distribution obtained in a routine experiment by the author. (Histogram is based on 2520 scores; smooth curve is normal distribution with same mean, variance and area as histogram).

1. History

Although nonparametric statistics can be traced as far back as 1710, when John Arbuthnott attempted to prove the wisdom of Divine Providence using the statistical Sign test, the preponderance of such tests are of quite recent origin. Van Dantzig and Hemelrijk (7) distinguish four stages of statistical development. In the first or one-parameter stage statistical quantities were considered to be constants such as the ratio of the yearly number of deaths to number of living. In the second or two-parameter stage variability was recognized as a factor and it was believed that empirical distributions could be described by stating the mean and variance, the parent distribution being assumed to be a normal distribution. In the third or multiparameter stage, universal normality was no longer an article of faith, but it was believed that an empirical distribution could be described by identifying its moments in the assumption that "statistical phenomena were governed by laws of general validity albeit that they showed somewhat greater complexity than just the normal law." The various Types of Pearsonian Curve were a product of this phase. In the fourth or no-parameter phase efforts to identify parameters of a parent population in order to be able to specify its probability law were largely replaced by attempts to determine "exact relations, valid for restricted sample sizes." Savage (38) places the "true beginning" of nonparametric statistics in 1936, and it is indeed at about this time that it began to take the form of a separate statistical discipline. The rapid growth of activity in this field since that date can be inferred from Figure 2 which shows the proportion of

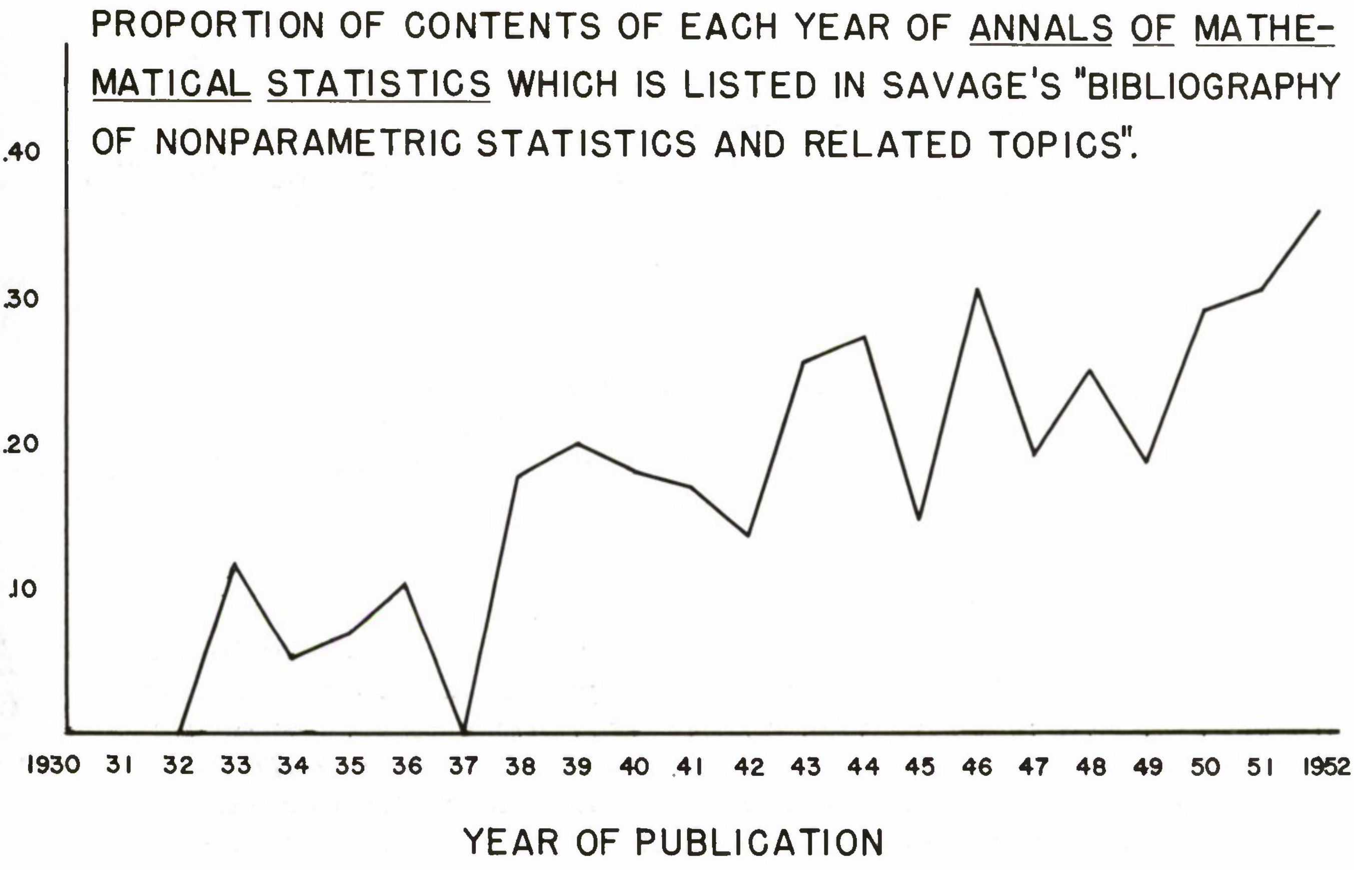


Figure 2. Twenty year growth of activity in the area of nonparametric statistics (as broadly defined by Savage).

articles in each volume of the Annals of Mathematical Statistics which are listed in Savage's "Bibliography of Nonparametric Statistics and Related Topics".

2. Definitions

The terms "nonparametric" and "distribution-free" are neither semantically satisfactory nor synonymous. This matter has been discussed at length by Kendall and Sundrum (28) who have attempted definitions of the terms which reflect the theoretical limitations of the tests to which they are commonly applied. Popular usage, however, has equated the terms and they will be used interchangeably throughout this report. Grossly speaking, a nonparametric test is one which makes no hypothesis about the value of a parameter in a statistical density function, while a distribution-free test is one which makes no assumptions about the precise form of the sampled population. Frequently the assumption is made that it is continuously distributed and sometimes more elaborate assumptions are made such as the assumption that the sampled populations have identical shapes or distributions symmetrical about the same point. However, the assumptions are never so elaborate as to imply a population whose distribution is completely specified. The term distribution-free is somewhat deceptive, however. The reason that no elaborate assumptions are made about the distribution of population magnitudes is very simple: the magnitudes are not used as such in the test. Instead, the ranks, ordinal position, frequency or some such attribute of the original observations provide the "information" used by the test statistic. And of course the "population" distribution of the attribute used must be known exactly for the conditions stated in the null hypothesis, just as must the population distribution of magnitudes in classical statistical tests. An important distinction should be made, however. While both parametric and nonparametric tests require that the form of a distribution be fully known, that knowledge, in the parametric case, is generally not forthcoming and the required distribution of magnitudes must therefore be "assumed" or inferred on the basis of approximate or incomplete information. In the nonparametric case, on the other hand, the distribution of the attribute is usually known precisely from a priori considerations and need not, therefore, be "assumed." The difference, then, is not one of requirement but rather of what is required and of certainty that the requirement will be met.

Because they do not use magnitudes as such, distribution-free tests do not test for parameters computed from them in the same sense that classical tests test for equal means, say, or identical variances. Instead, the analogous distribution-free tests might test for equal medians or identical interquartile ranges, i.e., values which can be computed from nonmagnitudinal attributes such as frequency, or position in rank order. Of course, a distribution-free test may be indirectly a test for parameters based on magnitudes; for example, if symmetrical populations can be assumed, then a distribution-free test for equal medians becomes, in addition, a test for equal means.

Although distribution-free tests generally are not based directly upon the magnitudes of the original observations, results by Stuart (46, 47) suggest that inferences from some such tests may be extended to the original magnitudes with a high degree of approximation. Stuart found very high correlations between observations, from either the normal or theuniform distribution, and their ranks. The correlations were respectively .94 and .96 for samples of 25 observations, and increased with increasing sample size toward limits of .98 and 1.00. The existence of these correlations is dependent merely upon the existence of a variance.

3. Distribution-Free vs Classical Tests

Both distribution-free and classical tests have points of superiority, and which type of test should be used depends upon a number of specific conditions as well as upon the sophistication of the user. The comparison, however, is generally quite favorable to distribution-free tests. Some advantages and disadvantages of distribution-free relative to parametric tests are outlined in the paragraphs to follow.

a. Simplicity of Derivation. Most distribution-free tests can be derived using simple combinatorial formulae, while the derivation of classical tests requires a level of mathematics far above the highest level attained by the typical research worker. However, the logic and appropriateness of a test's application, the assumptions it makes, and its sensitivity to assumption violation all hinge upon its derivation. If the research worker understands the derivation, he can deduce or infer much of this necessary information for almost any application he may contemplate, thus operating with a maximum of comprehension and flexibility. If he does not understand it, he is reduced to the uncomprehending "cookbook" procedures of performing tests by following a paradigm while obeying certain highly overgeneralized rules of thumb.

In the opinion of the writer this simplicity of derivation is by far the most important advantage of distribution-free statistics since, for research workers ignorant of higher mathematics, it replaces a mystery-cloaked ritual with a truly scientific procedure.

b. Ease of Application. The mathematical operations required in computing the test statistic are generally much less involved for distribution-free than for parametric statistics. Frequently all that is required is counting, or adding, subtracting and ranking. This simplicity of application is obviously an economic advantage, permitting lower-paid, mathematically naive personnel to be employed to reduce data and perform computations.

c. Speed of Application. When samples are of small or moderate size, distribution-free methods are generally faster than parametric techniques. This saving in computation time may be used to obtain more data, thus frequently cancelling any advantage the parametric test may have in terms of statistical efficiency. When samples are large (say $N \geqq 30$) distribution-free tests involving simple counting are generally faster, while those involving ranking may prove considerably more time consuming, than standard classical tests. And if a large number of similar tests are to be performed using an electronic computer, rather than a desk calculator, parametric tests are probably faster at all sample sizes.

d. Statistical Efficiency. As indicated in the preceding paragraphs, when judged by the practical criterion of the total amount of human effort required to conduct an experiment and analyze its results, distribution-free tests are frequently, if not generally, more efficient than their parametric counterparts. When judged by the mathematical criterion of statistical efficiency, distribution-free tests are often superior or equal to their most efficient parametric counterparts when both tests are applied under "nonparametric" conditions, i.e., conditions meeting all assumptions of the distribution-free test, but failing to meet some of the assumptions of the parametric test. When both tests are applied under "parametric" conditions, i.e., conditions meeting all assumptions of the parametric test, and therefore of both tests, distribution-free tests are very slightly less efficient (i.e., have relative efficiencies a shade less than 1.00) at extremely small sample sizes, becoming increasingly less efficient as sample size increases. When sample size becomes infinite, distribution-free tests generally have their lowest efficiencies relative to the most efficient, comparable parametric test. This efficiency value may be as high as .955 or as low as zero, depending on the test.

e. Scope of Application. Because they are based on fewer and less elaborate assumptions than classical tests, distribution-free tests can be legitimately applied to a much larger class of populations.

f. Susceptibility to Violation of Assumptions. Obviously the more elaborate the assumptions the fewer the number of situations which meet them, and, in this sense, parametric assumptions are the more susceptible to violation. For example, the parametric assumption of normality requires that, in addition to being continuously and symmetrically distributed (as might be assumed by nonparametric tests), the population must also be bell-shaped, since these are all features of a Gaussian distribution.

g. Detectability of Violations of Assumptions. When the nonparametric assumption of continuous distributions is violated, both the fact and the degree of the violation are readily apparent from the existence of tied scores in the obtained data. No such obvious indication advises the experimenter that a parametric assumption has been violated. Of course he may apply tests for normality or homogeneity to the obtained data, but such tests are rather unsatisfactory. They are unlikely to detect any but the most extreme violations when samples are small, and they are almost certain to detect even the most trivially slight violations when samples are very large.

h. Effect of Assumption Violations.* Although much has been written about the robustness of classical tests and their insensitivity to violation of assumptions, this claim actually rests upon a multitude of qualifications which rarely accompany it. The writer has obtained completely natural and uncontrived experimental data which, by violating a single parametric assumption, rendered a standard parametric

*This topic is discussed at length in two WADC Technical Reports shortly to go to press: Bradley, J. V., Studies in research methodology. I: Compatability of psychological measurements with parametric assumptions., and Bradley, J. V., Studies in research methodology II: Consequences of violating parametric assumptions - fact and fallacy.

test completely powerless, at reasonable sample sizes and standard significance levels, to reject any of a wide range of false hypotheses. The fact is that any violation of assumptions can be expected to alter the distribution of the test statistic and change the value at which the test statistic becomes significant. Whether or not this effect is negligible depends not only upon the degree to which the assumption is violated but also upon extrinsic factors such as sample size and significance level. This is true of both parametric and distribution-free tests.

In the nonparametric case, the effects of violation of the continuity assumption can be mitigated by applying certain methods of dealing with tied scores; in the parametric case, the effect of non-normality can be reduced by use of transformations, but at considerably greater expenditure of time.

i. Type of Measurements Required. Measurements on an interval or ratio scale are generally required by classical tests. However, distribution-free tests have greater versatility. They generally require measurements on at least an ordinal, or sometimes a nominal, scale but can be used with measurements from any higher order scale. They are, of course, the only truly appropriate tests when original scores exist in the natural form of ranks or small frequencies.

j. Logical Validity of Rejection Region. The distribution of a classical test statistic is usually continuous, increasing or decreasing smoothly, without fluctuation, except for a possible change of direction at a single mode. Unfortunately the point probability of a nonparametric test statistic does not necessarily always increase as the test statistic approaches its most probable value. It may level off or even dip before resuming its climb. This characteristic, when it exists, may be decidedly embarrassing when the rejection region for a distribution-free test is selected, on an intuitive basis. Should the rejection region be chosen as the cumulative probability for those values of the test statistic, which are least likely, or those which are most distant from the expected value of the test statistic?

k. Types of Statistics Testable. Statistics defined in terms of arithmetical operations upon observation magnitudes can be tested by classical techniques, while those defined by order relationships (rank) or category-frequencies can be tested by distribution-free methods.

Means and variances are examples of the former, medians and exceedances of the latter. The two approaches are different, but neither is superior; both types of statistic have their advantages.

l. Testability of Higher Order Interactions. Higher order interactions can be tested with ease by classical methods. However, there are few distribution-free tests for higher interactions and they are awkward and limited in application.

m. Choice of Significance Level. The distribution of the test statistic, when the null hypothesis is true, is usually continuous for classical tests and discrete for distribution-free tests. This means that, for any designated significance level α, a value of the classical statistic can be found whose cumulative probability is exactly α while, for the distribution-free test, such a value of the test statistic usually does not exist. Thus when using a classical test the research worker may choose any significance level he wishes, while, when using a distribution-free test, he must either accept one of the discrete cumulative probabilities of the test statistic as his significance level, or he must apply the test inexactly, using as significance level a cumulative probability which the test statistic cannot actually assume and rejecting whenever it is found to have a smaller cumulative probability. The latter choice is often forced upon him by inexact tables of probabilities which list values of the test statistic which are "significant" at the standard significance levels, .05, .01 and .001.

n. Influence of Sample Size. The size of the sample upon which they are to be used is an extremely important factor in determining the relative merits of distribution-free and classical tests. When samples are small (say $N \leq 10$) distribution-free tests are easier, quicker and only slightly less efficient even if all assumptions of the parametric test have been met. At these sample sizes, violations of parametric assumptions generally have their most devastating effect, yet are most unlikely to be detected. Therefore, unless the experimenter has a priori knowledge that all parametric assumptions have been met, the wiser choice would generally appear to be a distribution-free test. When samples are large (say $N > 30$); some distribution-free tests still compare favorably with their parametric counterparts. Others, however, will have become more laborious and time consuming, and, in contrast to parametric tests whose assumptions are met, their calculated or tabled probabilities may be only approximate. Finally, their efficiency relative to a parametric test whose assumptions are all true may have dropped to an appreciably low level. On the other hand, appreciable violations of parametric assumptions will have become more readily detectable and, in many cases, their effect may have become negligible due to the effect described by the central limit theorem. At large sample sizes, therefore, either type of test may be superior; however, circumstances are much more favorable to parametric tests than is the case when samples are small.

4. Organization of Material

Certain topics appear to be of critical importance to the understanding and application of distribution-free tests. These topics will be discussed in a general way in the following paragraphs and the same topics will form the paragraph headings under which each of the more important distribution-free tests will be examined.

a. Rationale. The best insurance against misapplication is a thorough understanding of the derivation and the mathematical logic upon which a test is based. The hypothesis which can be tested, the assumptions which must be made, the seriousness of various degrees of assumption-violation, the best method of dealing with such violations, the efficiency of the test, the situations to which it is applicable and the exactitude of the tables or of the probabilities obtained by formula all depend upon the test's derivation and can either be directly determined or partially inferred from a knowledge of it. Furthermore, many tests are legitimately applicable in situations for which they were not originally designed; however, the experimenter will not be able to recognize these situations unless he understands the derivation. Because of their importance, therefore, derivations have been given at some length. An effort has been made to use the simplest mathematics possible and to present derivations which will give the greatest insight into the logic of application and the advantages and limitations of the test. For this reason, many of the derivations are mathematically inefficient and are not in the form in which they are found in the literature.

b. Null Hypothesis. The literature on a test frequently does not contain an explicit and precise statement of the tested hypothesis. Instead the hypothesis may be implicit in some mathematical manipulations, it may be vaguely hinted at, or it may be stated explicitly but inaccurately, generally in the direction of overstatement. A major reason for these difficulties appears to be the lack of concise verbal terms to express what the test is actually doing. In order to avoid misleading the reader, an attempt has been made to express the tested hypothesis explicitly and precisely, with resort to expression in mathematical terms when necessary.

c. Assumptions. Assumptions also are frequently unstated, and occasionally misstated, in the literature, in which case they must be inferred from the derivation. In common with parametric tests, the assumptions of random sampling and independent observations are usually required. These assumptions however refer, at least in a sense, not to characteristics of the sampled population but rather to the method of sampling. Unlike "population" assumptions, their validity can generally be assured by adhering rigidly to certain prescribed sampling and experimental procedures.

Aside from the above one of the commonest nonparametric assumptions is that the sampled populations are continuously distributed. Such a population has an infinite number of abscissae and thus contains an infinite number of different score magnitudes, each of which has zero a priori probability of being drawn. Theoretically, therefore, a sample from a continuously distributed population will contain no scores of zero and no tied scores since zero is a predesignated score and since the first-drawn member of a tied group can be considered to predesignate the remainder. Zero scores are embarrassing in tests using the algebraic sign of scores, and tied scores are undesirable in tests which rank scores and whose derivation requires that each rank occur only once. The assumption of continuity, however, is an unrealistic one. Even if the sampled population is continuous, measurements made upon its members must be discretely distributed since no measuring instrument is capable of infinite precision. Suppose any population of actual measurements to be transformed into measurements on a scale running from zero to one and that precision is possible out to the N-th decimal place. Then the population of measurements is a discrete population whose interval width is the difference between successive digits at the N-th decimal place. The assumption of continuous distributions, therefore, can never be exactly fulfilled in practice. It can be approximated by taking fine measurements from distributions representing a very large number of distinguishable values. Fortunately, the degree to which the continuity assumption is violated can be largely inferred from the proportion of tied scores in the data. Therefore, although unrealistic, this assumption has the advantage that its violations are highly detectable.

Another assumption frequently encountered is that the sampled populations have identical, but unspecified, shapes. This assumption is found in tests which fail to reject when the sampled populations are identical but which may reject for a variety of reasons. By assuming identical shapes, rejection may be attributed to nonidentity of location. It is to be noted that this assumption may be dispensed with if the test be regarded merely as a test for identical populations against the broad alternative of nonidentical populations.

d. Treatment of Zero or Tied Scores. As mentioned earlier some tests require that all scores have an algebraic sign, i.e., that there are no scores of zero magnitude; others require that no scores have the same magnitude, i.e., that there are no ties for any given rank. Zero and tied scores do sometimes occur, however, and several methods of dealing with them have been suggested:

(1) Randomize. Randomly assign a plus or a minus to each zero score (say, on the basis of a coin toss); or randomly assign to scores of the same magnitude the ranks they would have if not tied, i.e., if differing very slightly. This method appeals to mathematicians, because only under this method does the test statistic have exactly the same distribution, when the null hypothesis is true, that it would have if the continuity assumption were not violated. It makes little sense experimentally, however, since it permits an additional and, in a sense, unnecessary, element of pure chance to help determine whether or not a false hypothesis will be rejected.

(2) Minimize the Probability of Rejection. Assign all zero scores that algebraic sign which is least conducive to rejection of the null hypothesis; or assign ranks to tied scores in the way least conducive to rejection of the null hypothesis. This is the conservative approach and it alone insures, in advance of sampling, that the tested hypothesis will not be falsely rejected due to violation of the assumption of continuity.

(3) Obtain the Average Value of the Test Statistic. Assign half the zeros a plus, half a minus sign; or assign each score in the tied group the average of the ranks the members of the group would have if not tied. The latter is known as the midrank method. It results in a distribution of ranks having the same mean but somewhat smaller variance than the discrete rectangular distribution of integers 1 to N. For some tests a "correction for ties" has been devised for use with

the midrank method. When applied to asymptotic formulae for the test statistic the correction compensates for the reduction in variance due to the use of midranks. It thus tends to reestablish the validity of the test in the large-sample case. The logic of the implicit assumptions upon which this correction is based has been challenged. (VII-36) However, the correction is probably an improvement in any case, although perhaps not fully restoring the test to exactitude.

(4) Obtain the Average Probability. Break ties in all possible ways, calculate the test statistic and obtain its probability for each way, and average these probabilities. This improves on the above method by obtaining the average probability of the test statistic, rather than the probability for the average value of the test statistic, averaging over all possible ways in which tied measurements could have been caused by truly differing scores. It is time consuming, however, and has the disadvantage, in common with the preceding method, that the average of all possibilities may differ greatly from that one possibility which represents the true state of affairs.

(5) Drop Zeros. Discard zero scores and reduce N accordingly. The power of certain tests has been found to be greater under this method than under methods (1) or (3). However, it seems likely that this is an artifact attributable to an unrecognized and spurious increase in the probability of rejection in all cases, i. e., when the tested hypothesis is true as well as when it is false. Zero difference scores lend support to the hypothesis of "no difference." Discarding them eliminates data favoring the null hypothesis and permits contrary data to assume greater weight, thus spuriously increasing the probability of rejection.

A final method is to calculate the test statistic twice, once giving all ambiguous data (zero or tied scores) the possible true values which are most conducive to rejection, once giving them the values least conducive to rejection. It has been said with some justification, that if in both cases the test statistic falls within, or in both cases outside of, the rejection region there is no problem; if it does not, there is no solution.

e. Efficiency. Certain mathematical properties of a test are important in evaluating its usefulness. The power of a test is the probability of its rejecting a specified false hypothesis. (It is equal to $1-\beta$ where β is the probability of committing a Type II error - failing to reject a false null hypothesis.) Power, then, depends upon at least four variables: (a) the amount by which the hypothesis is in error, i.e., the size of the discrepancy, δ, between the hypothesized and true condition, (b) the size, α, of the significance level chosen, (c) the location of the rejection region, e.g., whether the test is one-tailed or two-tailed, (d) the size, N, of the sample used in the test. A power function is a curve in which all but one of these variables are held constant and power is plotted as ordinate against that one variable, usually δ, as abscissa. A test of a given true hypothesis is most powerful against a specified alternative hypothesis if no other test of the same hypothesis has greater power against the same alternative. If it is most powerful with respect to each member of a class of alternative hypotheses, the test is called uniformly most powerful against that class of alternatives.

A test is unbiassed, for a given alternative, if the probability of rejecting the null hypothesis is greater when the alternative hypothesis is true than when the null hypothesis is true.

A test is consistent for a given alternative to the null hypothesis if, when that alternative hypothesis is true, the probability of rejecting the false null hypothesis, i.e., the power of the test, approaches 1 as the sample size, N, on which the test is based, approaches infinity. The test is consistent with respect to a class of alternatives if it is consistent for each of the alternatives of which the class is composed.

Efficiency is a relative term comparing the sensitivity of a test with that of some other test, usually the most powerful alternative available. Let A and B be statistical tests of the same null hypothesis against the same set of alternative hypotheses, and let the tests use the same significance level and the same number of tails. Then the efficiency of test A relative to test B can be interpreted as the ratio b/a, where a is the number of observations required by test A to equal, by some criterion, the power of test B based on b observations. There are actually a number of definitions of efficiency, differing mainly in the criterion by which the two powers are equated.

Asymptotic efficiency is usually defined in terms of the limiting value of the ratio b/a as b approaches infinity and is therefore relevant only when the test is to be applied to very large samples. It has the advantage of being very nearly independent of the exact size of the samples so long as they are very large. The more common definitions of asymptotic efficiency appear to be equivalent. Asymptotic relative efficiency, abbreviated A. R. E., and sometimes called Pittman efficiency, is defined roughly as follows. Let A and B be two consistent tests based upona and b observations respectively, each test statistic being asymptotically normally distributed. Let both A and B test a null hypothesis H_o against an alternative hypothesis H_a at a significance level α. The asymptotic relative efficiency of A with respect to B is the limiting value of the ratio b/a as a is allowed to vary in such a way as to give A the same power as B while, simultaneously, b approaches infinity and H_a approaches H_o. The purpose of the "approach" of H_a to H_o is to prevent the ratio b/a from assuming a limiting value of 1 which it otherwise would do since at extremely large sample sizes the power of a consistent test against a fixed alternative is virtually 1. The method of obtaining asymptotic relative efficiency has been shown to be equivalent (Stuart V-50) to that of obtaining asymptotic local efficiency. Let A and B be one-tailed tests based on a and b observations respectively and testing the same null hypothesis against the same set of alternative hypotheses at the same significance level. Let b approach infinity and vary a so that the power functions of the two tests have equal slopes at the point H_o. Then the limiting ratio b/a is the asymptotic local efficiency of test A relative to test B. Somewhat similar methods involve taking the asymptotic ratio of first derivatives, i. e. slopes, of the power functions at the point H_o. In the case of equal-tailed, two-tailed tests this is zero and the asymptotic ratio of second derivatives is used. Estimate efficiency is obtained by establishing a mathematical equivalence between relative efficiency of two tests and the relative efficiency of two estimators of a population parameter. The latter requires that both estimates be consistent and asymptotically normally distributed and is expressed in terms of the ratio of the asymptotic variances of the two estimators. Estimate efficiency is therefore an index of relative efficiency for the case where both tests are based upon large, i. e. "infinite", samples. Stuart (VI-26) observes that estimate efficiency is equivalent to asymptotic relative efficiency. All of the asymptotic efficiencies defined above refer to the relative power of two tests at the point H_o of their power functions. The efficiency values obtained therefore represent the effectiveness of one test relative to another when the true condition differs negligibly from the hypothesized condition, i. e., when the alternative hypothesis lies in the immediate vicinity of the null hypothesis.

Nonasymptotic efficiencies depend upon the size sample upon which the test is based, upon the location of the rejection region, upon the size α of the significance level chosen, and upon the alternative hypothesis or set of alternative hypotheses. Balancing the disadvantage that nonasymptotic efficiencies are highly specific to experimental test conditions, is the advantage that they are quite realistic to those conditions. While asymptotic efficiencies provide a limiting value for a test's efficiency at infinite sample size, this value is generally much lower, when distribution-free statistics are compared with classical tests, than is the efficiency value at practical sample sizes. The relative efficiency of A with respect to B is simply b/a where a is the number of observations required by test A to equal the power of test B based on b observations when both statistics test the same null hypothesis against the same alternative at the same significance level (both either one-tailed or two-tailed). The power efficiency of test A with respect to test B (of the same null hypothesis at the same significance level against the same set of alternative hypotheses) is obtained by holding a constant and varying b until the power functions of the two tests are equated in the sense that the area between the power functions when the ordinate for test A exceeds that of test B equals the area between the power functions when the reverse is true. The value taken by b need not be integral. The power efficiency of A relative to B is then b/a. This definition of efficiency has the advantage that the obtained efficiency values are peculiar to an entire class of alternative hypotheses rather than to a specific alternative hypothesis. Its disadvantage lies in the failure of statisticians to agree completely upon the precise method by which to apply it.

Some asymptotic efficiencies of some distribution-free tests relative to their classical counterparts are given in Table I. All efficiencies given in the body of the table are for the case where both tests are applied under conditions satisfying all of the assumptions of the classical test. Except when otherwise specified, the tests were applied to normally distributed populations; comparisons involving Student's t required that the two populations to which both tests were applied have equal variances, etc. When more than one efficiency is listed in a cell, the asymptotic efficiency of the test depends upon the number of categories or groups to which the test is applied. An asymptotic efficiency of zero requires some interpretation. It means that, when both tests are based upon an equal and "infinite" number of observations, the test with zero asymptotic efficiency requires "infinitely" more observations in order to equal the power of the comparison test. It does not mean that the ratio of the powers of the two tests is zero or infinity. The power of any consistent test

TABLE I

EFFICIENCIES OF SOME DISTRIBUTION-FREE TESTS RELATIVE TO, AND UNDER CONDITIONS ASSUMED BY, A (MOST POWERFUL) CLASSICAL, COMPARISON STATISTIC*

Test	Asymptotic Efficiency	Established by	Footnotes
Student's t*	1.000		
X-test	1.000	van der Waerden	
Mann-Whitney	.955	Pitman, Mood, Dwass, van der Waerden	C, U, 1
Sign	.637	Cochran, Jeeves & Richards, Dixon, Walsh	C
Westenberg Median	.637	Mood	
No. Runs (Location)	0	Pitman, Mood	C
Analysis of Variance*	1.000		
Kruskal-Wallis H	.955	Andrews	C, 2
Friedman	.637-.912	Friedman	
k-Sample Median	.637	Andrews	C, 3
F - Ratio*	1.000		
Mood's Dispersion	.87	Mood, Dwass	
No. Runs (Dispersion)	0	Pitman, Mood	C
Maximum Likelihood*	1.000		
S_1 for Dispersion	.74	Cox & Stuart	
S_3 for Dispersion	.71	Cox & Stuart	
Correlation Coeff.*	1.000		
Kendall' τ	.912	Moran	
Spearman's ρ	.912	Hotelling & Pabst	
Blomqvist's Median Test	.405	Blomqvist	
Regression Coeff. b*	1.000		
Mann's T	.985	Stuart	C, U
Daniels	.985	Stuart	
Cox & Stuart's S_1	.860	Stuart	
Cox & Stuart's S_3	.827	Stuart	
Cox & Stuart's S_2	.782	Stuart	
Median test for Trend	.782	Stuart	
Rank Serial R_h	0	Stuart	C
Records test d	0	Stuart	C
Difference sign	0	Stuart	C
Turning Point	0	Stuart	

C - test has been shown to be consistent under certain conditions.
U - test has been shown to be unbiased under certain conditions.
1 - Asymptotic efficiency is 1.000 when populations have uniform distributions (Pitman).
2 - Asymptotic efficiency is 1.000 when populations have uniform distributions (Andrews).
3 - Asymptotic efficiency is .333 when populations have uniform distributions (Andrews).

TABLE II

POWER COMPARISONS OF SOME STATISTICAL TESTS APPLIED TO THE SAME DATA

Tests in Order of Decreasing Power (within a block)	Null Hypothesis	Assumptions	Sample Sizes	Sig. Level	Author and Type of Comparison
Student's t-test X-test Mann-Whitney Max.Absolute Deviation Number of Runs	Equal Means	Normal Distributions Equal Variances	3,3; 1,∞; 2,∞; 5,∞; 3,7 3,7; 5,6 3,3; 3,7; 5,6; 1,∞; 2,∞ 3,7; 5,6; 5,∞ 3,7; 5,6; 5,∞	.05	van der Waerden Mathematical
X-test Mann-Whitney Student's t-test	Equal Means	Uniform Distributions Equal Variances	4,6	.05	van der Waerden Mathematical
Mann-Whitney Max. Absolute Deviation Westenberg Median	Equal Means	Normal Distributions Equal Variances	5,5	.025	Dixon Mathematical
Mann-Whitney Tsao's Max. Abs. Dev. Epstein's Exceedances Number of Runs	Equal Means	Normal Distributions Equal Variances	10, 10	.05	Epstein Empirical
Lehmann's Most Powerful Mann-Whitney (1-tailed) Westenberg Median " Mann-Whitney (2-tailed) Westenberg Median " Max. Absolute Deviation Number of Runs	Identical Populations against y's Distributed as Maximum x's	Continuous Distributions	4,4; 6,6	.10	Lehmann Mathematical
Regression Coefficient b Mann's T-test Daniels Foster & Stuart's D Foster & Stuart's d Rank Serial Correlation Difference Sign Turning Point	Randomness against Linear Trend	Normal Distributions	100	.05 & .01	Foster & Stuart Empirical
Number of Runs Longest Run	Randomness vs. Markoff Chain			.05	Bateman Mathematical

approaches 1 as sample size approaches infinity. Therefore when a consistent test has an asymptotic efficiency of zero both its power and the power of the comparison test are very close to 1 and are approaching 1 as sample size approaches infinity. The power of the comparison test, however, is approaching 1 faster. That is, at any "infinite", i.e. extremely large, sample size the power of the comparison statistic is very slightly greater than that of the test whose efficiency is sought, but "infinitely", i.e. very many, more observations are required by the test with zero asymptotic efficiency to close this infinitesimal power gap. Finally, tests with zero asymptotic efficiency with respect to the same comparison test do not necessarily have equal asymptotic efficiency with respect to one another. For example, each of the four tests in Table I having zero asymptotic efficiency with regard to the regression coefficient has zero asymptotic efficiency with respect to all of the seven to ten tests listed above it.

A number of investigators have compared the relative powers of distribution-free tests with respect to each other without actually calculating small-sample efficiencies. They have simply been compared under identical conditions of application and then ranked in order of power. Sometimes a most powerful classical statistic was included. The results (see Table II) of these comparisons are naturally highly peculiar to the conditions under which the comparison occurred.

Certain statisticians (17, 31, 49, 50) have addressed themselves to the problem of determining "most powerful" distribution-free tests. Although successful, the gain in power is usually slight and is generally obtained at the expense of simplicity. Furthermore, the property of greatest power is contingent upon the type of distribution assumed to exist when the null hypothesis is false. Lehmann (31) has obtained the most powerful rank test for the hypothesis that two populations have idrntical distributions against the alternative that the second population is distributed as the k largest observations in the first population. Terry (49) has described the rank test which is asymptotically most powerful, at the point H_0, for testing the hypothesis of identical distributions against the alternative that the two populations are normally distributed with the sare variance but with different means. His test procedure requires that the $N_1 + N_2$ observations be ranked in order of magnitude irrespective of sample. He then substitutes for each rank the average magnitude corresponding to that rank in the average sample of size $N_1 + N_2$ from a normal distribution with zero mean and unit variance. This is accomplished by means of tables (XX and XXI) supplied by Fisher and Yates (13). Thus scores from a population of unknown form are, in a sense, transformed so as to represent scores from a normal distribution. Exact tables of probabilities are available for Terry's test for $N_1 + N_2 \leqq 10$, an asymptotically normally distributed test statistic being used for large samples. A somewhat similar test, the X-test, has been proposed by van der Waerden (50, 51). The power of the X-test can equal that of Student's t-test when applied to normally distributed populations (50) and can exceed the power of the t-test when both are applied to uniformly distributed populations (52). Both Terry's and van der Waerden's tests are analogous to, and appear to be slightly more powerful than, the Mann-Whitney test. Both have the dubious advantage of giving greater "weight" to extreme observations than does the Mann-Whitney test (7). Neither, however, can compare with the latter in simplicity or ease of application. Furthermore the quality of high power against "parametric", i. e. normal, alternatives, while useful is not an overriding consideration in selecting a nonparametric test. It is a useful property in those cases where populations are normal and variances homogeneous but the experimenter does not have certain knowledge of this fact, i. e., when a distribution-free test is necessitated by the experimenter's ignorance rather than the population's nonnormality.

f. Application. The applicability of most tests is directly deducible from the derivation as is the method of application. Furthermore, many, if not all, distribution-free tests are applicable in situations other than those for which they were originally designed, and it would be quite impossible to anticipate all such situations and to outline the test's method of application in each of them. Therefore, only the briefest example will be given of the application of each distribution-free test, and the "Application" section will often be used to illustrate or expand upon points made in presenting the test's derivation.

g. Discussion. Tests which upon superficial examination appear to be quite distinct may actually be identical or similar in function, i.e., may ultimately perform the same or nearly the same mathematical operation. In other cases, although different, they may be mathematically interrelated to a high degree. Not infrequently the author of a test overstates, understates or misstates the test's capabilities. Such matters are taken up in each test's "Discussion" section.

h. Tables. For most distribution-free tests probabilities are based upon simple combinatorial formulae. The point probability of a given value of the test statistic is generally a fraction whose numerator is the number of different ways (combinations) in which that value of the test statistic can be obtained and whose denominator is the sum of the number of different ways in which all possible values of the test statistic can be obtained. Such tests are usually exact for small samples whose N is small enough to permit enumeration of the combinations constituting the numerator of the (cumulated) probability fraction. (The denominator is usually easy to obtain.) The time and labor involved in these computations increases drastically with increasing N, however, so that exact tables frequently do not extend beyond an N of very moderate size. For larger N's approximate probabilities may generally be obtained fairly easily from asymptotic formulae, and at this point the tables, if they continue, become inexact. The approximation is usually very good for large values of N. There is sometimes a gap, however, between the largest N for which exact probabilities have been tabled and the smallest N at which the asymptotic approximation is good.

The existence of adequate tables is an important criterion for the acceptability of a distribution-free test. There is practically no limit to the number of distribution-free tests which can be devised on a sound mathematical basis. However, a test for which no tables have been computed is of very limited value unless exact cumulated probabilities can be easily computed by formula, or unless the asymptotic approximation is good at small sample sizes, neither of which is likely to be the case.

i. Sources. The survey of literature upon which this report is based was confined almost entirely to publications written in English. However, not all of the relevant English publications were reviewed and only a fraction ot those reviewed are reported. The number of relevant articles is immense and increases exponentially as one broadens one's definition of what is nonparametric. An attempt was made only to cover tests, of broad applicability, whose probabilities can be calculated exactly when samples are small, and which, when sampling from a continuously distributed population, do not specify the exact form of that distribution. This criterion, for example, eliminated tests of card matching, which apparently find application only in experiments on extra-sensory perception, approximate tests or parametric tests used in violation of their assumptions, and tests requiring such nonclassical but specific distributions as a Poisson or an exponential. Despite efforts at thoroughness, however, it is virtually certain that relevant tests meeting all these criteria have escaped the writer's attention; in some cases such tests were detected, but were unobtainable. No claim is made for complete coverage; however, it is felt that a core of better known and more important tests has been covered fairly adequately.

In the following chapters tests have been grouped together largely on the basis of a common type of mathematical derivation, sometimes according to the type of sample information used, and occasionally according to the type of function which the test serves. Only the simplest, most extensively tabled, and most promising tests have been treated at length. Sources are referenced in the treatment of each test and are listed at the end of each chapter. (Occasionally reference will be made to a source listed in the bibliography of a different chapter, in which case the Arabic reference number will be preceded by a Roman numeral indicating the number of the chapter in which the referenced source is listed.) Because the number of sources relevant to a given test or to a general topic may be quite large, those sources regarded as most critical have been indicated by printing their authors' names in capital letters. Primary sources (or, in some cases, the nearest thing to a primary source) for a unique distribution-free test have been indicated by an asterisk. Sources containing tables of probabilities for a distribution-free test have been indicated by placing a capital T in the left margin. If the source contains tables for more than one such test, two T's are used; and, if a table is an extensive one, the T is underlined.

BIBLIOGRAPHY

1. Bahadur, R. R. and Savage, L. J., The nonexistence of certain statistical procedures in nonparametric problems. Annals of Mathematical Statistics, 1956, 27, 1115-1122.

2. BLUM, J. R.. and FATTU, N. A., Nonparametric methods., Review of Educational Research, 1954, 24, 467-487.

3. Bradley, R. A., Some notes on the theory and application of rank order statistics. Parts I and II., Industrial Quality Control, 1955, 11.

TT 4. Burington, R. S. and May, D. C., Handbook of probability and statistics with tables, Handbook Publishers Inc., Sandusky, Ohio, 1953.

5. Chernoff, H., A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics, 1952, 23, 493-507.

6. Craig, C. C., Recent advances in mathematical statistics, II. Annals of Mathematical Statistics, 1942, 13, 74-85.

7. VAN DANTZIG, D. and HEMELRIJK, J., Statistical methods based on few assumptions. (also errata and addenda), Bulletin of the International Statistical Institute, 1954, 34.

TT 8. Dixon, W. J. and Massey, F. J., Introduction to statistical analysis, New York: McGraw-Hill, 1951, pp. 247-263.

9. Dwass, M., On the asymptotic normality of certain rank order statistics. Annals of Mathematical Statistics, 1953, 24, 303-306.

10. Dwass, M., On the asymptotic normality of some statistics used in non-parametric tests, Annals of Mathematical Statistics, 1955, 26, 334-339.

TT 11. Edwards, A. L., Statistical methods for the behavioral sciences, New York: Rinehart, 1954, pp. 181-212, 399-439.

12. Fisher, R. A., *Contributions to mathematical statistics*, New York: Wiley, 1950.

TT 13. Fisher, R. A. and Yates F., *Statistical tables for biological agricultural and medical research*, 3rd Ed., New York: Hafner, 1949.

14. Fraser, D. A. S., *Nonparametric methods in statistics*, New York: Wiley, 1957.

T 15. Hald, A., *Statistical theory with engineering applications*, New York: Wiley, 1952.

16. Hoeffding, W., A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 1948, 19, 293-325.

17. Hoeffding, W., "Optimum" nonparametric tests. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, 1951, pp. 83-92.

18. Hoeffding, W., Some powerful rank order tests. (abstract) *Annals of Mathematical Statistics*, 1952, 23, 303.

19. Hoeffding, W., The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 1952, 23, 169-192.

20. Hoeffding, W. and Rosenblatt, J., The efficiency of tests. *Annals of Mathematical Statistics*, 1955, 26, 52-63.

T 21. Hoel, P. G., *Introduction to mathematical statistics*, 2nd Ed., New York: Wiley, 1954, pp. 281-303.

22. Hotelling, H. and Pabst, Margaret R., Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, 1936, 7, 29-43.

TT 23. Isaac, W., *Some nonparametric tests and their tables*, (mimeographed) 45 pp.

24. James, G. and James, R. C., *Mathematics dictionary*, New York: van Nostrand, 1949.

TT 25. Kendall, M. G., Rank correlation methods, 2nd Ed., New York: Hafner, 1955.

26. Kendall, M. G., The advanced theory of statistics, Vol. II, London: Griffin, 1946.

27. KENDALL, M. G. and BUCKLAND, W. R., A dictionary of statistical terms, New York: Hafner, 1957.

28. KENDALL, M. G. and SUNDRUM, R. M., Distribution-free methods and order properties. Review of the International Statistical Institute, 1953, 3, 124-134.

*T 29. Kruskall, W. H. and Wallis, W. A., Use of ranks on one-criterion variance analysis. Journal of the American Statistical Association, 1952, 47, 583-621.

* 30. Lehmann, E. L., Consistency and unbiasedness of certain nonparametric tests. Annals of Mathematical Statistics, 1951, 22, 165-179.

* 31. Lehmann, E. L., The power of rank tests. Annals of Mathematical Statistics, 1953, 24, 23-43.

32. Lehmann, E. L. and Stein, C., On the theory of some non-parametric hypotheses. Annals of Mathematical Statistics, 1949, 20, 28-45.

* 33. Mood, A. M., Introduction to the theory of statistics, New York: McGraw-Hill, 1950, pp. 385-418.

34. Moses, L. E., Nonparametric methods, Chapter 8 in Walker, Helen and Lev, J., Statistical inference, New York: Holt, 1953, pp. 426-450.

35. Moses, L. E., Non-parametric statistics for psychological research. Psychological Bulletin, 1952, 49, 122-143.

36. Noether, G. E., On a theorem by Pitman. Annals of Mathematical Statistics, 1955, 26, 64-68.

37. Ruist, E., Comparison of tests for non-parametric hypotheses. Arkiv für Matematik, 1955, 3, 133-163.

38. SAVAGE, I.R., Bibliography of nonparametric statistics and related topics. Journal of the American Statistical Association, 1953, 48, 844-906.

39. SAVAGE, I. R., Contributions to the theory of rank order statistics: two sample case,I. Annals of Mathematical Statistics, 1956, 27, 590-615.

40. Savage, I. R., Nonparametric statistics. Journal of the American Statistical Association, 1957, 52, 331-344.

41. Scheffé, H., Statistical inference in the non-parametric case. Annals of Mathematical Statistics, 1943, 14, 305-332.

42. SIEGEL, S., Nonparametric statistics. American Statistician, 1957, 11, 13-19.

TT 43. SIEGEL, S., Nonparametric statistics for the behavioral sciences, New York: McGraw-Hill, 1956.

44. Smith, K., Distribution-free statistical methods and the concept of power efficiency. In Festinger, L. and Katz, D. (Eds.) Research methods in the behavioral sciences, New York: Dryden Press, 1953, pp. 536-577.

45. Stevens, S. S., On the theory of scales of measurement. Science, 1946, 103, 677-680.

46. Stuart,A., The correlation between variate-values and ranks in samples from a continuous distribution. British Journal of Statistical Psychology, 1954, 7, 37-44.

47. Stuart, A., The correlation between variate-values and ranks in samples from distributions having no variance. British Journal of Statistical Psychology, 1955, 8, 25-27.

48. Stuart, A., The cumulants of the first n natural numbers. Biometrika, 1950, 37, 446.

*T 49. Terry, M. E., Some rank order tests which are most powerful against specific parametric alternatives. Annals of Mathematical Statistics, 1952, 23, 346-366.

* 50. van der Waerden, B. L., Order tests for the two-sample problem and their power. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1952, 55, 453-458.

51. van der Waerden, B. L., Order tests for the two-sample problem and their power. (Corrigenda) Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1953, 56, 80.

52. van der Waerden, B. L., Order tests for the two-sample problem II and III. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1953, 56, 303-310, 311-316.

53. Wald, A. and Wolfowitz, J., Statistical tests based on permutations of the observations. Annals of Mathematical Statistics, 1944, 15, 358-372.

54. Wallis, W. A., Rough-and-ready statistical tests. Industrial Quality Control, 1952, 8, 35-40.

55. Whitworth, W. A., Choice and chance. New York: Hafner, 1951.

*TT 56. Wilcoxon, F., Some rapid approximate statistical procedures. American Cynamid Co., Pamphlet, 1949.

57. Wilks, S. S., Mathematical statistics, Princeton, N. J: Princeton University Press, 1950.

58. WILKS, S. S., Order statistics. Bulletin of the American Mathematical Society, 1948, 54, 6-50.

TT 59. Wilson, E. B., An introduction to scientific research, New York: McGraw-Hill, 1952, pp. 185-189, 197-202, 229-231, 247-250, 266-268.

60. Wolfowitz, J., Additive partition functions and a class of statistical hypotheses. Annals of Mathematical Statistics, 1942, 13, 247-279.

61. Wolfowitz, J., Non-parametric statistical inference. Proceedings Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, 1949, pp. 93-113.

62. Yule, G. U. and Kendall, M. G., An introduction to the theory of statistics, New York: Hafner, 1950, pp. 19-68, 187-189, 258-272, 459-481.

CHAPTER II

TESTS BASED ON THE BINOMIAL DISTRIBUTION

A number of distribution-free test statistics are binomially distributed. They are among the simplest, safest, most nearly exact and most extensively tabled nonparametric tests. Their statistical efficiency is not the highest, but is generally not so low as to nullify their other advantages. The sample information used by most of them is simply the direction of the difference between two scores, i. e., the algebraic sign of the difference. Binomial tests are extremely versatile, finding application in testing for location, trend (in either location or dispersion), randomness of predicted order, and in the setting of confidence limits for quantiles.

1. Introduction

Suppose that all of the possible outcomes of an event may be dichotomized into two mutually exclusive categories, arbitrarily labeled "success" and "failure", these two outcomes having probabilities p and $q = 1 - p$ respectively. Then if the event is permitted to occur n times, the probability that r of the n outcomes will be successes is $P_r(r) = \binom{n}{r} p^r q^{n-r}$ which is the general expression for a term in the expansion of the binomial $(p+q)^n$.

Proof: The probability that r successes and n - r failures will occur in a specified order is $p^r q^{n-r}$. For example, letting subscripts indicate order of appearance, the probability for the order in which all successes occur first, followed by all failures, is the product $(p_1)(p_2) \ldots (p_r)(q_{r+1})(q_{r+2}) \ldots (q_n) = p^r q^{n-r}$. However, since we seek only the probability of a given <u>frequency</u> of successes, the probability $p^r q^{n-r}$ of a given frequency of successes occurring in a specified pattern must be multiplied by the number of patterns which r successes and (n-r) failures can assume. If the n units (p's and q's) were all distinguishable, the number of unique patterns would be n!, the number of permutations of n things. They are not all distinguishable however. In each distinguishable pattern, the r successes can be permuted with one another in r! ways without changing the pattern. And for each such permutation of successes, the n-r failures can be permuted in (n-r)! ways without changing the appearance of the pattern. The number of permutations, n!, then must be the number of distinguishable patterns times r!(n-r)!, the number of ways each distinguishable pattern can be permuted without altering its appearance. The number of distinguishable patterns is therefore $\frac{n!}{r!(n-r)!}$, which is, of course, the number of combinations of n things taken r at a time, frequently expressed by the symbol $\binom{n}{r}$. The probability of exactly r successes in n trials is therefore

$\binom{n}{r} p^r q^{n-r}$, and the cumulative probability, i. e., the probability of r or fewer successes in n trials is $\sum_{i=0}^{r} \binom{n}{i} p^i q^{n-i}$.

The binomial term $\binom{n}{r} p^r q^{n-r}$ expresses the probability for r successes out of n trials only if the following conditions, implicit in its derivation, are met:

(a) Outcomes must be capable of being *dichotomized* (Since only two outcome probabilities, p and q, are used in the derivation.)

(b) The two outcome categories must be *mutually exclusive* (since $q = 1 - p$).

(c) The outcome of the n events must be completely *independent.* (Since the same value, p, is used to express the probability of success on each of the n trials, the probability of success on a single trial must not change from one trial to another and, therefore, must not be influenced by the outcome of any other trial.)

(d) "Events" must be *randomly selected.* (The formula $\binom{n}{r} p^r q^{n-r}$ gives the probability that by *chance* r successes will occur in n trials if the *chance* probability of success in a single trial is p. If events are not randomly selected, then outcomes are susceptible to nonchance influences.) There must therefore be no bias or system in the selection of which n trials, out of an infinite population of potential trials, to test. Specifically, among other things this means that none of the valid data may be systematically excluded from the test.

The above qualifications will appear in modified form as *assumptions* for all tests whose test statistic is binomially distributed. Such tests are outstanding among distribution-free tests for two reasons: First they are extremely simple, both in derivation and in application. Second exact probabilities for both the point (20, 28) and cumulative (34, 25, 28) binomial have been extensively tabled. Thus, while for most distribution-free tests large n's require probabilities to be calculated approximately from asymptotic formulae, in the case of binomial tests exact probabilities are readily attainable for many large samples.

The mean and variance of a binomially distributed variate are np and npq respectively (for proof see Hoel [I-21] pp. 65-67), and when n is large and p is close to .50 the binomial is closely approximated by the normal distribution. The critical ratio for r, the number of successes, is therefore $\frac{|r-np| - 1/2}{\sqrt{npq}}$, the 1/2 being a correction for continuity. The normal approximation should not be used except for those cases not covered by the extensive binomial tables which are now available. The approximation is reasonably good so long as the product np is greater than 5. Even when this criterion is met, however, the approximation is likely to be poor at the extreme tails of the distribution, especially when n is small (say less than 100). The inaccuracy of the normal approximation can be expected to increase therefore with decreasing n, with increasing departures of p from .50 in either direction, and with decreasing, i.e. more and more extreme, significance levels.

2. The Sign Test for the Median Difference

a. Rationale. Suppose that n pairs of measurements have been taken, one member of each pair having been taken under condition A, the other under condition B, and that a B measurement is as likely to exceed as to be exceeded by its paired A measurement. Then, if zero differences are impossible, the differences $A_i - B_i$ can be either positive or negative and the outcome "positive" is binomially distributed with probability $p = 1/2$. For example, John Arbuthnott (1) found that every year from 1629 to 1710 the number of males born in the city of London exceeded the number of females. If male and female babies are equally likely, the chance probability of the reported results is

$\sum_{i=0}^{o} \binom{n}{i} (1/2)^n = (1/2)^{82}$. (Arbuthnott obtained this result and interpreted the excess of male births as a manifestation of Divine Providence, which he believed to be allowing precisely for the greater mortality rate among males "who must seek their Food with danger", so as to leave a perfect equality of sexes at the age of mating.)

b. Null Hypothesis. For every $A_i - B_i$ difference, $P_r(A_i > B_i) = P_r(A_i < B_i) = 1/2$. Sufficient conditions for its validity are that both the A population and the B population are continuously distributed and the population of A - B differences has a median of zero.

c. Assumptions. Since binomial tests require that outcomes must be of two types only, there must be no zero differences, i.e., the members of no pair shall be "tied." Frequently this requirement is expressed by the more restrictive assumption that the population of differences is continuously distributed. Since the outcomes of binomial events must be independent, the sign of the difference for one pair must have no influence upon the sign of the difference for any other pair. This means among other things, that a given A measurement shall be paired once and only once with a measurement from the B population. Finally, the sample of measurements must have been randomly selected from the parent population of differences.

d. Treatment of Ties. The null hypothesis is that

$P_r(A_i > B_i) = P_r(A_i < B_i) = 1/2$. Therefore $P_r(A_i = B_i)$ must equal zero. Zero differences constitute a third category of outcomes. Since the Sign test is based upon the binomial distribution which requires that outcomes fall into two mutually exclusive classes, zero differences are decidedly embarrassing. They can occur for two reasons: because a noninfinitesimal proportion of the parent population of differences is zero, or because, although this is not the case, zero differences are obtained due to the inability of the measuring instrument to achieve infinite precision. In the former case, the Sign test simply is not appropriate. For the latter case, various methods have been recommended for disposing of zero differences. They can be dropped and n reduced accordingly (14, I-8, 27). Half may be treated as plusses, half as minuses (8, 27). They may be replaced by signs "drawn" randomly from an infinite population half of whose members are plusses, half of which are minuses (27). Or all zeros may be treated as if they had the algebraic sign least conducive to rejection of the null hypothesis.

The Sign test has greatest power when zero differences are dealt with according to the first alternative. However, the greater power resulting from use of this method is not necessarily an argument for its adoption. A zero, being in a sense "halfway between" a plus and a minus suggests that plusses and minuses are equally likely. By ignoring, i.e. discarding, data which lend support to the null hypothesis, one naturally increases the probability of rejecting that hypothesis and consequently enhances the power of the test. The probability of rejecting a true null hypothesis has also increased, however, and the apparent gain in power is attributable to a subtle increase in the "true", as contrasted with the nominal, significance level. For example, consider 1000 differences of which 960 are zero, 13 plus and 27 minus. If half of the zeros are regarded as plus and half as minus and the two-tailed Sign test is applied to the 493 plusses and 507 minuses, the cumulative probability is .681. If the zero differences are discarded and the test is applied to the 13 plusses and 27 minuses, the cumulative probability falls within the .05 level of significance. Assuming that half the zeros actually represent plus scores, half minus scores, the "true" cumulative probability is .681 in both cases. However, in the latter case the experimenter believes his significance level to be .05 when actually the true significance level corresponding to this alleged figure would be some figure greater than .681. Thus discarding the zeros biases the test toward rejection.

The "randomization" method preserves exactly the mathematical conditions upon which the validity of the Sign test depends. However, it makes little sense experimentally. Normally one interprets small chance probabilities as implying the presence of a nonchance effect. But if it is known that pure chance determined a substantial portion of one's results, then small chance probabilities may imply unlikely chance effects as strongly as (or more strongly than) nonchance effects. In such cases the null hypothesis may remain as reasonable as any alternative hypothesis. Ambiguities may also arise in marginal situations. Suppose for example that an experimenter using the .05 level of significance obtains significant results after "randomizing" zeros, but discovers that his results would have a "chance" probability of .15 had he regarded half the zeros as plusses, half as minuses. The reverse situation would be equally distressing.

The first three methods of dealing with zero differences are based upon an implicit assumption that zero differences represent true differences which, if measured with infinite accuracy, would be found to be positive half of the time, negative half of the time. However, if zero differences are due to imprecisiond measurement, as it is assumed, such a 50-50 split is by no means assured. The "measuring instrument" might be such that all differences between -.0015 and +.0040 were measured as zero. One would then expect the preponderance of recorded zeros to represent true plusses.

None of the methods of dealing with zero differences, therefore, is entirely satisfactory. Giving all zeros the sign least conducive to rejection is the safest method while, in the long run, the average probability error is minimized by treating half the zeros as plusses, half as minuses. If only a small proportion of the differences are zero, say less than 5%, one would expect the "error" introduced by zero differences generally to be of small practical consequence. However, when zeros constitute a substantial proportion of the data, considerable caution should be used in applying the Sign test.

e. Efficiency. A normal distribution is symmetrical with median equal to mean. Therefore, if applied to a normally distributed population of differences, the Sign test for the median difference is equally a test for the mean difference and can legitimately be compared with Student's t-test. Under the conditions stated, the one-tailed Sign test has, relative to Student's t, an asymptotic efficiency of $2/\pi$ or .637. This same figure is obtained whether the asymptotic efficiency be an estimate efficiency (4, 44) A. R. E. (15), or an efficiency of certain other types (7, 15, 16). It refers, of course, to the case where the discrepancy δ between the true difference and hypothesized difference is zero, i.e., very slight. If samples are of infinite size, the efficiency of the Sign test is independent of the size α of the significance level, but decreases from .637 to a limiting value of .500 as δ increases from zero to infinity (15).

The small sample efficiency of the Sign test depends strongly upon the precise definition of efficiency chosen (2). It decreases with increasing values of n, α and δ (7). Small sample efficiencies as high as .96 have been found (43).

Power functions for the Sign test have been published by Dixon (7) and by Walsh (42). Stewart (36) has prepared tables giving the sample size at which a false null hypothesis ($p = .50$) will have a given probability of rejection, i.e., test will have a given power, at the .05 level of significance, for various "true" values of p. The test is consistent provided only that $p \neq q$, i.e., in the present case provided only that the null hypothesis is false (14).

f. Application. Subtract each B score from its matched, i.e. paired, A score. If a small proportion of the differences are zero, "assign" half of them a positive sign, half a negative sign; if there are an odd number of zero differences, discard one zero difference, reduce n by one, and proceed as above. Let r be the number of plusses and n-r be the number of minuses after the zeros have been "assigned." Then the cumulative probability of obtaining r or fewer plusses by chance if the null hypothesis is true is $\sum_{i=0}^{r} \binom{n}{i} 1/2^n$.

If a two-tailed test is required, one rejects the null hypothesis if this cumulative probability equals or is less than $\alpha/2$ or equals or exceeds $1-\alpha/2$. If a one-tailed test is required and the alternative hypothesis is that the median difference is less than zero, the null hypothesis is rejected if $\sum_{i=0}^{r} \binom{n}{i} 1/2^n \leq \alpha$. For the opposite alternative, reject if the summation equals or exceeds $1-\alpha$.

g. *Tables.* Probabilities can be most accurately obtained from tables of the cumulative binomial (34, 25, 28, 46) entered with p=.50. Other tables (4, 8, 26, I-8, I-23, I-43, I-59) have been designed specifically for the Sign test.

h. *Discussion.* Mathematically the Sign test simply tests the hypothesis that the parameter, p, of a binomial population has the value .50. In equivalent experimental terms it tests the null hypothesis that the population of A-B differences has a median of zero. The inference is frequently made that if the median difference is zero, then the A population and the B population are equally "good" in a quantitative sense. Such an inference cannot legitimately be made without introducing an additional assumption: that the A-B differences are *symmetrically* distributed about zero. Without this assumption one can legitimately infer that half of the units comprising the A population are superior to the units with which they happen to be matched in the B population and that half of the B units are superior to their paired mates from the A population, but *not* that these two "superiorities" represent equivalent difference magnitudes. It is to be noted that the assumption of symmetry requires that the *mean* difference be zero.

By adding M to each B score before subtraction from its paired A score, one can test the null hypothesis that the median difference is M. If the assumption of symmetry can justifiably be made, one can test the hypothesis that the mean difference is M, or, in other words, that the A population is on the average M units "better" than the B population. By multiplying each B score by 1+100p before subtraction, one can, under the assumption of symmetry, test the hypothesis that the A population is on the average p percent "better" than the B population. (See 8 or 26)

The preceding discussion has assumed that every A score has the same parent population and likewise for every B score. Actually the formula holds good even if every A or B score comes from a different population so long as each population corresponding to a given A-B difference has zero median. The null hypothesis tested is that all of the populations from which the A-B differences were "drawn" have zero median. This type of application should be approached with caution, however. Suppose, for example, that half of the pairs represent populations in which A's are truly superior to B's while the reverse is true for the other half. Although the null hypothesis is entirely false, the probability of its rejection is no greater than if it were true. Again, suppose that for a tenth of the pairs A's are truly superior to B's while for the remainder there is no real difference. The power of the test would be much greater if that tenth of the data were tested separately. Applications of the type described, therefore, may greatly reduce the power of the test, and even when the null hypothesis is rejected, it is not at all clear what alternative hypothesis is indicated. Finally, in this type of application, the modifications described in the preceding paragraph become meaningless and should not be used.

It has been stated that the Sign test is particularly appropriate when the members of each pair were subjected to similar treatment, but when treatments differed from one pair to another. This, of course, represents a special case of the application discussed above. Here it is implied that a number of variables may have a real effect upon the absolute values of the A's, the B's or even the A-B differences, but that only one variable, the one in which the experimenter is interested, can have a real effect upon the direction of the A-B differences, i.e., the signs of the differences. This is not necessarily an unrealistic assumption. For example, the A's and B's might be positions of seismograph needles during, and an hour previous to, an hypothesized tremor. The seismographs being located in widely different parts of the world, the A-B differences would be expected to vary in size with distance from the source of tremor. Furthermore, the numerical size of the difference might be reported in metric units by some and in British units of measurement by others. These considerations would preclude the use of a t-test, but not the Sign test since the variable mentioned would affect the size but not the direction of the differences.

It is extremely important, however, that no variable causing differences between pairs shall interact with the variable in which the experimenter is interested, i. e., shall differentially affect the sign of the difference between members of a pair. Suppose, for example, that A and B are two strains of wheat and that some of the AB pairs were grown in a northeastern county, the rest in a southwestern county. If the former location has a moist climate, the latter a dry one, it may well be that A is superior to B in one location and inferior in the other. Subjecting pairs to different treatments, therefore, may introduce subtle and spurious interactions between "tested" and "nontested" effects with the result that the power of the test is reduced and the true alternative hypothesis may differ greatly from the alleged one.

i. Sources. 1, 2, 4, 7, 8, 10, 12, 13, 14, 15, 16, 26, 27, 30, 36, 42, 43, 44, 45, I-2, I-3, I-8, I-11, I-21, I-23, I-28, I-35, I-43, I-54, I-59.

3. The Sign Test for the Median

a. Rationale. Suppose that n observations, X_i's, are taken from a continuously distributed population whose median is M. Then half of the observations, on the average, should fall above M, half below, i. e., the number of observations falling above M is binomially distributed with p= .50. Thus, the number of observations above an hypothesized median M can be used to test the validity of the hypothesis. But the number of observations above M is the same as the number of positive differences if M is subtracted from each observation. The Sign test for the median, therefore, is equivalent to the Sign test for the median difference in which the X_i's constitute the A population and the B population consists of the single value M.

b. Null Hypothesis. For every X_i, $P_r(X_i > M)=P_r(X_i < M)=1/2$.

Sufficient conditions for its validity are that the X's are drawn independently and are continuously distributed with a common population median

M. It is in fact only necessary to assume that the X's are continuously distributed in the neighborhood of M.

c. Assumptions. (1) $P_r(X_i = M) = 0$, i.e., none of the observations must fall on the hypothesized median.

(2) Whether a given X_i falls above or below M is independent of the position of any other X_i with respect to M. This implies among other things that either the population is an infinite one, which will be the case if it is continuously distributed, or sampling is with replacement.

(3) The X_i's must have been randomly selected from their respective populations.

d. Treatment of Observations Falling on the Hypothesized Median. See 2. Treatment is analogous.

e. Efficiency. See 2. Efficiencies quoted under 2 apply with equal validity to the test for the median.

f. Application. Count the number, r, of X's which are smaller than M. If a small proportion of the X's equal M, count half of them as smaller than M. If there are an odd number of such tied X's, discard one of them and reduce n by 1. For a two-tailed test at the level α, reject the null hypothesis if

$\sum_{i=0}^{r} \binom{n}{i} 1/2^n \leq \alpha/2$ or $\geq 1 - \alpha/2$. If the alternative hypothesis for a one-tailed test is that the population median exceeds M, reject the null hypothesis if $\sum_{i=0}^{r} \binom{n}{i} 1/2^n \leq \alpha$. For the opposite one-tailed alternative hypothesis, reject if the summation $\geq 1 - \alpha$.

g. Tables. See 2 and the paragraph below.

h. Discussion. If the X's are arranged in order of increasing magnitude with subscripts indicating rank in that order (1 = smallest, n = largest), then if r observations are below M, M exceeds the value X_r, i.e., $M > X_r$. Therefore, rejecting the null hypothesis because r observations have fallen below the median is equivalent to rejecting it because the median exceeds X_r. Walsh (43) has prepared tables of probabilities for the Sign test for the median which call for this approach.

If the X's all come from the <u>same</u> continuously distributed population whose mean equals its median (which will be the case if the population is symmetrically distributed), the Sign test for the median is equivalent to a test for the mean. In other words at the cost of introducing two new assumptions, homogeneous populations and symmetrical distribution, the Sign test for the median becomes a Sign test for the mean. By adding (or subtracting) a constant C to every X before applying the test, the hypothesis can be tested that the population mean has "slipped" a distance C below (or above) a value it is known to have had at some earlier period.

i. <u>Sources.</u> See 2.

4. <u>Cox and Stuart's</u> S_2 <u>Sign Test for Trend in Location</u>

a. <u>Rationale.</u> Suppose that 2n measurements have been recorded or are available in an order of sequence and it is desired to test whether the sequence may contain a monotonic, i.e., nonreversing, trend. If there is no trend of any kind, i.e., if sequential position has no effect upon measurement magnitudes, these magnitudes will be randomly distributed in sequence. If measurements are divided into independent pairs and if in each pair the measurement later in sequence is subtracted from the earlier measurement, the sign of each difference will be as likely to be plus as to be minus. If zero differences are impossible, the number of differences of one sign will be binomially distributed. On the other hand, if a unidirectional trend exists differences of one sign will tend to predominate.

b. <u>Null Hypothesis.</u> Let subscripts represent the position of a given measurement in the sequence of 2n measurements. The null hypothesis, then, is that for every

$$X_i \text{ with } i \leq n \text{ the } P_r\,(X_i > X_{i+n}) = P_r\,(X_i < X_{i+n}) = 1/2.$$

Sufficient conditions for its validity are that the X's are continuously distributed and are randomly related to sequence, i.e., contain no trend.

c. Assumptions. (1) $P_r(X_i = X_{i+n}) = 0$ for every $i \leq n$, i.e., the members of no pair are tied.

(2) Whether a given X_i falls above or below X_{i+n} is independent of the outcome for any other pair.

(3) The X's are randomly selected.

d. Treatment of Ties. The authors recommend counting half the zero differences as plusses, half as minuses. Also see 2..

e. Efficiency Applied to populations known to be normally distributed, the S_2 test for trend in location has asymptotic relative efficiency .78 with respect to the best parametric test, based on the regression coefficient (37). Under the same conditions, it has A. R. E. .79 compared to Spearman's or Kendall's rank correlation tests used as tests of randomness (5). For other comparisons, see Table I.

f. Application. If the total number of measurements is not an even number, drop the middle measurement to make it so. Let 2n stand for the number of measurements remaining. From each X_i in the first half of the sequence, subtract the corresponding measurement X_{i+n} in the second half. If a small proportion of the differences are zero, assign half of them a plus, half a minus. If an odd zero remains, discard it and reduce n by 1. Let r be the number of positive differences. Then for a two-tailed test at significance level α reject the null hypothesis if $\sum_{x=0}^{r} \binom{n}{x} 1/2^n$ either equals or is less than $\alpha/2$ or equals or exceeds $1-\alpha/2$. For a one-tailed test at the level α reject the null hypothesis if $\sum_{x=0}^{r} \binom{n}{x} 1/2^n \leq \alpha$ if alternative hypothesis is an upward trend (or $\geq 1-\alpha$ if alternative hypothesis is a downward trend).

g. Tables. See 2 .

h. Sources. (5, 11)

5. Cox and Stuart's S_3 Sign Test for Trend in Location

a. Rationale. See 4, substituting "3n" for "2n".

b. Null Hypothesis. Let subscripts represent the position of a given measurement in the sequence of 3n measurements. The null hypothesis, then, is that for every

$$X_i \text{ with } i \leq n \text{ the } P_r (X_i > X_{i+2n}) = P_r (X_i < X_{i+2n}) = 1/2.$$

Sufficient conditions for its validity are that the X's are continuously distributed and are randomly related to sequence, i.e., contain no trend.

c. Assumptions. See 4, substituting "X_{i+2n}" for "X_{i+n}".

d. Treatment of Ties. See 4,

e. Efficiency. Applied to populations known to be normally distributed, the S_3 test for trend in location has A. R. E. .83 with respect to the best parametric test, based on the regression coefficient (37). Under the same conditions, it has A. R. E. .84 compared to Spearman's or Kendall's rank correlation tests used as tests of randomness (5). For other comparisons see Table I.

f. Application. If the total number of measurements is not divisible by 3, "add" one or two "dummy" measurements in the middle of the sequence to make it so. Let 3n stand for the number of measurements as modified. From each X_i in the first third of the sequence, subtract the corresponding measurement X_{i+2n} in the last third. The data in the middle third will not be used. If a small proportion of the differences are zero, assign half of them a plus, half a minus. If an odd zero remains, discard it and reduce n by 1. Let r be the number of positive differences. Then for a two-tailed test at significance level α, reject the null hypothesis if $\sum_{x=0}^{r} \binom{n}{x} 1/2^n$ either equals or is less than $\alpha/2$ or equals or exceeds $1-\alpha/2$. For a one-tailed test at level α, reject the null hypothesis if $\sum_{x=0}^{r} \binom{n}{x} 1/2^n \leq \alpha$ if alternative hypothesis is an upward trend (or $\geq 1-\alpha$ if alternative hypothesis is a downward trend).

g. Tables. See 2.

h. Discussion. The S_3 test uses only 2/3 of the raw data employed by the S_2 test; however, the members of each pair of measurements whose difference is taken are 1/3 farther apart. The net result is an increase in efficiency. If a real trend exists, then the farther removed two measurements are in sequence, the greater the expected difference in magnitudes and the more likely that the sign of the difference will betray the direction of the trend. The S_2 test, however, has one advantage. Since it uses all of the data, statistical inference can be extended to the entire parent population. Strictly speaking, inferences based on the S_3 test cannot legitimately be extended to the middle third of the sampled sequence, since a temporary trend occupying only this portion could not be detected.

i. Sources. 5, 11, 37.

6. Cox and Stuart's S_3 Sign Test for Trend in Dispersion

a. Rationale. Suppose that 3kn measurements have been recorded in order of sequence and it is desired to test whether the dispersion of the measurements about a linear regression line changes monotonically with position of measurements in the sequence. If the true dispersion remains constant, then the ranges of consecutive sets of k measurements should vary on a chance basis only. And if the range of a subsequent set is subtracted from that of an earlier set, the difference is as likely to be positive as to be negative. If zero differences are impossible, the number of differences of one sign will be binomially distributed. On the other hand, if dispersion changes monotonically with position in sequence, differences of one sign will tend to predominate.

b. Null Hypothesis. Let w_i represent the range of the i th consecutive set of k measurements. The null hypothesis, then, is

that for every w_i with $i \leq n$ the $P_r(w_i > w_{i+2n}) = P_r(w_i < w_{i+2n}) = 1/2.$

Sufficient conditions for its validity are that the X's are continuously distributed with constant dispersion about a linear regression line.

c. Assumptions. (1) $P_r(w_i = w_{i+2n}) = 0$ for every $i \leq n$, i.e., the members of no pair are tied. If the X's are continuously distributed, the w's will be also and the assumption will be satisfied.

(2) Whether a given w_i falls above or below w_{i+2n} is unaffected by the outcome for any other such pair.

(3) The X's are randomly selected.

d. Treatment of Ties. See 4.

e. Efficiency. Applied to populations known to be normally distributed, the S_3 test for dispersion has A.R.E. of .71 compared to the maximum likelihood test (5).

f. Application. The selection of the integer k is arbitrary and will not affect the validity of the test; however, it can be expected to affect the test's power. Letting N stand for the total number of measurements, the following rule is suggested by the authors:

take $k = 2$ if $N < 48$, take $k = 3$ if $48 \leq N < 64$, take $k = 4$ if $64 \leq N < 90$,

take $k = 5$ if $N \geq 90$. Let n be the integral part of N/3k and drop N-3kn measurements from the middle of the sequence. Divide the 3kn remaining measurements into 3n consecutive sets of k measurements each. Find the range of measurements within each of the 3n sets. Finally, using these ranges as scores or measurements, proceed exactly as in the S_3 test for trend in location.

g. Tables. See 2.

h. Discussion. This test can be made a test for trend in variance, (or standard deviation) simply by substituting this term for "range" and applying the test as outlined above.

The authors do not suggest the use of the S_2 test to test for dispersion, although it obviously could be legitimately used for that purpose.

i. Sources. 5, 11.

7. Noether's Sequential Test for Linear Trend

Cox and Stuart's tests for trend in location give specific values to a constant, C, in a more general test discussed by Noether (23, 24). The latter author, in effect, sets the null hypothesis that

$P_r (X_i > X_{i+C}) = P_r(X_i < X_{i+C}) = 1/2$ and examines the optimum value of C for a sequential probability ratio test of that hypothesis.

8. Noether's Binomial Test for Cyclical Trend

a. Rationale. Suppose that 3n measurements have been recorded or are available in order of sequence and it is desired to know whether the sequence may contain a fluctuating or cyclical trend. If the measurements are continuously distributed and there is no trend of any kind, no two measurements will be equal, and the measurements will be randomly related to sequence. Any three consecutive measurements will be equally likely to have any of the six sequences represented by the six possible permutations of three things. However, of these six sequences only two are monotonic, i.e., ascend or descend without reversals, while the remaining four change direction in the middle. For example, if the three measurements are ranked, the ranks will be found to have one of the six sequences: 1 2 3, 3 2 1, 1 3 2, 2 3 1, 2 1 3, 3 1 2, the underlined sequences being monotonic. The probability of monotonicity for such a set of three measurements is therefore 1/3 if the sequence is random and the measurements are continuously distributed. And if the 3n measurements are divided into n independent sets of 3 consecutive measurements each, the number of monotonic sets will be binomially distributed with p = 1/3. On the other hand, if a cyclical or fluctuating trend of any but the shortest possible "wave length" exists, one would expect more than 1/3 of the sets to be monotonic.

b. Null Hypothesis. For every

$i \leq n$, the $P_r (X_{3i} > X_{3i-1} > X_{3i-2}) + P_r (X_{3i} < X_{3i-1} < X_{3i-2}) = 1/3.$

Sufficient conditions for its validity are that the X's are continuously distributed and the size of the X's is unrelated to their position in sequence.

c. Assumptions. (1) $P_r (X_{3i} = X_{3i+1}) = 0$ and $P_r (X_{3i+1} = X_{3i+2}) = 0$

for every $i \leq n$, i.e., adjacent scores in no set are tied.

(2) Whether or not any given set is monotonic is independent of the monotonicity or nonmonotonicity of any other set. Among other things, this means that no X is used in more than one set.

(3) The X's are randomly drawn.

d. Treatment of Ties. Ties are a practical problem only when the tied scores are members of the same set. If the first and third scores are tied and the second is not, the set is clearly nonmonotonic and there is no ambiguity. If adjacent members of a set are tied, the set is as likely as not to be monotonic; therefore, half of such sets should be counted as monotonic, half as nonmonotonic (the odd set, when it exists, being discarded and n reduced by 1). If all three members of a set are tied, the chance probability of monotonicity is obviously 1/3, and one third of such sets should be counted as monotonic (one or two sets being discarded and n reduced accordingly if the number of such sets is not divisible by 3).

e. Efficiency. Noether states that he does not believe the test to be highly efficient.

f. Application. If the total number of measurements is not divisible by 3, drop one or two measurements from the middle of the sequence to make it so. Let 3n stand for the number of measurements remaining. Divide these 3n measurements into n independent, i.e. nonoverlapping, sets of 3 consecutive measurements. Count the number of monotonic sets, treating tied members of a set as outlined above. Call this number r and call the total number of sets used n. Then for a one-tailed test at significance level α reject the null hypothesis if $\sum_{x=r}^{n} \binom{n}{x} (1/3)^x (2/3)^{n-x} \leq \alpha$. This tests H_0 against the one-sided alternative that once a direction is taken it tends to persevere for a longer than chance period. A two-sided test would include the alternative that direction fluctuates more rapidly than would be expected by chance. However, such a contingency seems unlikely to be of great practical interest, since such a fluctuation in this case would very nearly amount to alternation of direction, i.e., change with every measurement.

g. Tables. 34, 25, 28.

h. Discussion. This test is also presented by its author as a sequential probability ratio test.

Lehmann (17, 38) has briefly proposed a test of the hypothesis that two populations are identical, which is analogous to Noether's test. If 2n scores have been drawn from an X population, and 2n from a Y population, and if X's are paired at random with one another and then with a pair of likewise paired Y's, there will result n independent quadruples consisting of two X's and two Y's. If the null hypothesis is true, and the X's and Y's are continuously distributed, the chance probability that in a given quadruple both X's will either be greater than or less than both Y's is 1/3. The number of quadruples for which this is the case will therefore be binomially distributed with $p = 1/3$ and can be used to test the hypothesis of identical populations. The test is consistent if the sampled populations are continuous, ties are randomized and the alternative hypothesis is that $p \neq 1/3$.

i. Sources. 24.

9. Mosteller's Test of Predicted Order

a. Rationale. Suppose that n individuals each are to be tested under k conditions and the experimenter has reason to believe that he can predict the order of excellence of performance under the k conditions. If "performance" is continuously distributed so that no two conditions will result in the same score, then for any one individual there are k! orders in which the k conditions could be arranged. If performance is independent of the conditions under which it is tested, then each of the k! orders is equally likely with probability 1/k!. If performance is truly unrelated to differences among tested conditions, then the number of individuals whose order of performance has been correctly predicted is binomially distributed with $p = 1/k!$. On the other hand, if performance is related to conditions and if the experimenter has correctly predicted the relationship, the predicted order will tend to exceed its chance expectation.

b. Null Hypothesis. P_r (Order Predicted by Experimenter) $= 1/k!$. Sufficient conditions for its validity are that measurements are continuously distributed and unrelated.to the specific experimental conditions under which they occur.

c. Assumptions. (1) None of the performance scores for a single individual can be tied.

(2) The order of performance excellence for any given individual is unaffected by that of any other individual.

(3) Individuals and individual's scores are randomly selected.

d. Treatment of Ties. Ties are no practical problem unless one of the possible ways of "breaking" the ties results in the predicted order. In those cases, for every group of t tied scores, there will be t! ways of breaking the ties, and if there is more than one such group for a single individual, the number of ways of breaking the ties will be the product of these factorials. Therefore, for each individual whose order of performance contains ties and could be the predicted order if the ties are broken properly, find the number of ways in which ties could be broken. Sum these over all such individuals, and call the total "D". Let N stand for the number of such individuals. Then N/D is the proportion of these individuals whose order should be regarded as the predicted one, and (N/D) N or N^2/D individuals should be counted as having the predicted order. Simpler techniques, which err in the direction of conservatism, are to regard the N individuals as not having the predicted order, or to discard the N individuals and reduce n by N.

e. Efficiency. Apparently unknown.

f. Application. Treating ties by one of the techniques outlined above, count the number of individuals whose performance under the k conditions conforms exactly to the predicted pattern, i.e., whose performance excellence under each condition has the rank predicted for performance under that condition. Let this number be r and the total number of individuals tested be n. Since a smaller than chance number of individuals having the predicted order is unlikely to be of interest to the experimenter, only a one-tailed test of the opposite situation will be outlined. For a one-tailed test at the level α, reject the null hypothesis in favor of the alternative that the predicted order has a greater than "chance" probability if $\Sigma_{i=r}^{n} \binom{n}{i} (1/k!)^r (1 - 1/k!)^{n-r} \leq a$.

g. Tables. 34, 25, 28.

h. Discussion. It is very important to remember that this test tests only that if the k conditions affect performance differentially the experimenter has done a better than chance job of predicting the pattern. Suppose that of 15 conditions 10 affect performance in the same way and are therefore equivalent, while the remaining 5 conditions affect performance differentially. If the experimenter correctly assigns one of the ranks from 1 to 15 to each of the five differentiating conditions, the predicted rank order will occur more frequently than 1/15! of the time and the null hypothesis will tend to be rejected more than a of the time. However, the predicted rank order will not be correct for the 10 equivalent conditions since it will imply that they differ, which they do not. Suppose again that five conditions arranged in order of "excellence" are A B C D E and that the experimenter has predicted the order A B C E D. If the conditions differ greatly relative to performance variability, the experimenter's predicted order may be expected to occur less than 1/5! of the time; while, if performance variability is large relative to the true differences among conditions, the experimenter's predicted order may be expected to occur more than 1/5! of the time and the null hypothesis will tend to be rejected more than a of the time. The temptation to accept the predicted order as the correct one, when the null hypothesis is rejected, should therefore be resisted.

i. Sources. 34 (Introduction, pp. xxxvi-xxxvii).

10. Confidence Limits for Quantiles

a. Rationale. Assume that a random sample of n independent observations has been taken from an unknown but continuously distributed population, and that it is desired to establish confidence limits for the magnitude of a population quantile, Q. This quantile may be a percentile, quartile, median, or, more generally, that population magnitude below which some specified proportion p of the population lies.

Let the n sample observations be arranged in order of increasing magnitude with subscripts indicating rank position in that order, i.e., from smallest to largest the observations are X_1, X_2, X_3, ..., X_r, ..., X_s, ..., X_{n-2}, X_{n-1}, X_n. Also, let ϵ be an infinitesimally small positive magnitude. If Q lies at or below $X_r + \epsilon$ then r or fewer sample observations have fallen below the population quantile Q, the chance probability for which is $\sum_{i=0}^{r} \binom{n}{i} p^i (1-p)^{n-i}$, where p is the proportion of units in the population whose magnitude is less than Q. Likewise, if Q lies at or above $X_s - \epsilon$ then n - s +1 or fewer sample observations exceed Q, or equivalently, s - 1 or more observations are smaller than Q. The chance probability for this is $\sum_{i=s-1}^{n} \binom{n}{i} p^i (1-p)^{n-i}$.

With qualifications which will be outlined under "Assumptions", these two probabilities may be regarded as the probabilities that Q lies below $X_r + \epsilon$ and that Q lies above $X_s - \epsilon$ respectively. If s is larger than r, the events referred to by these two probabilities are mutually exclusive (since ϵ is an infinitesimal). Therefore the probability that Q is neither below $X_r + \epsilon$ nor above $X_s - \epsilon$ is

$$1 - \sum_{i=0}^{r} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=s-1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \text{ or } \sum_{i=r+1}^{s-2} \binom{n}{i} p^i (1-p)^{n-1}$$

and this is equivalently the probability that Q lies between $X_r + \epsilon$ and $X_s - \epsilon$ Since ϵ is an infinitesimal, it is also the probability that $X_r < Q < X_s$.

b. Assumptions. Random sampling and independent observations are assumed for reasons given in (1). The assumption of continuous distribution is required in order to rule out tied observations. Actually, ties become a practical problem only when they occur at the critical end points of the confidence region, i.e., when X_r is tied with X_{r+1} or X_s with X_{s-1}: Such ties render the end points of the confidence region indistinct and impose an additional (see next assumption) element of inexactitude upon the calculated confidence level. If X_r and X_{r+1} are tied, for example, then $X_r + \epsilon$ cannot be greater than X_r and equal to or less than X_{r+1} as required by the derivation. The tied observations X_r and X_{r+1} represent a third category of outcomes, e.g., on rather than above or below the median, thus rendering the binomial an inappropriate mathematical model. The assumption of a continuous distribution is also required because it implies an infinite population. If the population is infinite, the probability of an observation smaller than Q is p for every observation; if the population is finite, the probability for every observation after the first depends upon the outcomes of the previous drawings. The final assumption is incompatible with the immediately previous one. It is that there is zero probability that the population quantile Q lies between X_r and X_{r+1} or between X_{s-1} and X_s. The probability

that $X_r < Q < X_s$ was derived to be $\sum_{i=r+1}^{s-2} \binom{n}{i} p^i (1-p)^{n-i}$;

however, this is precisely the same probability which would have been

obtained for the event $X_{r+1} \leq Q \leq X_{s-1}$. But this implies that

$P_r (X_r < Q < X_{r+1}) = 0$ and that $P_r (X_{s-1} < Q < X_s) = 0$ which offends

common sense. Phrased differently, the derivation given under "Rationale" took ϵ to be an infinitessimal, but would have led to the same results if ϵ had been any positive value such that

$X_r + \epsilon < X_{r+1}$ and $X_{s-1} < X_s - \epsilon$. Again, this obviously implies the untenable assumption that Q cannot occupy the region between X_r and X_{r+1} or between X_{s-1} and X_s. The reason for the discrepancy is simply that "r observations below Q" and "r+1 observations below Q" are two "adjacent" eventualities in a discrete distribution of "number of sample observations below the population quantile Q". Since this is a discrete distribution of frequencies, there is no event "in between" the two named. However, "population quantile is X_r" and "population quantile is X_{r+1}" are nonadjacent eventualities in a continuous distribution of magnitudes assignable to the population quantile. An error has therefore been introduced by using a discrete distribution, i.e. the binomial, to express probabilities for a continuously distributed variable. In terms of confidence limits, the error is no larger than the difference between the confidence limits $X_r < Q < X_s$ and $X_{r+1} \leq Q \leq X_{s-1}$.

c. Treatment of Ties. If either X_r and X_{r+1} or X_{s-1} and X_s are tied, it is suggested that the confidence region be changed (i.e. shifted, expanded or contracted) so as to have untied endpoints. The conservative, i.e. safest, approach would be to reduce r or enlarge s to the extent necessary to include within the confidence region all observations which had been tied with the endpoints. The confidence level will, of course, have to be recalculated for the new confidence region determined by the reassigned values of r and s.

d. Application. Let Q be the unknown magnitude of the population score below which a specified proportion p of the population scores lie. Draw n sample observations from this population and rank these observations from smallest (1) to largest (n). Ties should be dealt with as outlined in the preceding paragraph. Take $\sum_{i=r+1}^{s-2} \binom{n}{i} p^i (1-p)^{n-i}$ to be the confidence level for the hypothesis that Q lies in one of the following confidence regions. If the most conservative probability statement is desired, take $X_r < Q < X_s$ as the confidence region. However, if greatest accuracy is desired in the sense of minimizing the error, take the confidence region to be $\frac{X_r+X_{r+1}}{2} \leq Q \leq \frac{X_{s-1}+X_s}{2}$. The former will usually be the more conscionable procedure. The values p, r, and s must, of course, be selected prior to sampling.

e. Tables. 34, 25, 28, See also 19, I-8 p. 360.

f. Discussion. The a priori probability that the magnitude of the r^{th} ranked observation will be less than Q is not the exact probability that the magnitude obtained for the r^{th} ranked observation will be less than Q. Even in the obtained sample, X_r could be assigned any magnitude between X_{r-1} and X_{r+1}, and still be the r^{th} observation in order of magnitude. The range of magnitudes "represented" by X_r, then, might be considered to be $\frac{X_{r-1}+X_r}{2}$ to $\frac{X_r+X_{r+1}}{2}$, i.e., the point halfway between X_r and the next lower magnitude to the halfway point to the next higher magnitude. Then if the rank of r represents magnitudes as high as $\frac{X_r+X_{r+1}}{2}$, the summation $\sum_{i=0}^{r} \binom{n}{i} p^i (1-p)^{n-i}$ would give the probability that Q lies below $\frac{X_r+X_{r+1}}{2}$ rather than below X_r. Obviously, then, the probability that Q is less than or equal to X_r is *no greater* than $\Sigma_{i=0}^{r} \binom{n}{i} p^i (1-p)^{n-i}$. Therefore, we can be confident *at least* at the level $\Sigma_{i=r+1}^{s-2} \binom{n}{i} p^i (1-p)^{n-i}$ that $X_r < Q < X_s$. By introducing an inequality then we can make a definitive probability statement which takes account of the error discussed under the last "assumption". It is that $P_r (X_r < Q < X_s) \geq \Sigma_{i=r+1}^{s-2} \binom{n}{i} p^i (1-p)^{n-i}$. Also, if instead of the most conservative probability statement, we wish to make the most nearly accurate one, we can take $\frac{X_r + X_{r+1}}{2} \leq Q \leq \frac{X_{s-1} + X_s}{2}$ as the most probably "true" confidence interval corresponding to the confidence level $\Sigma_{i=r+1}^{s-2} \binom{n}{i} p^i (1-p)^{n-i}$.

If Q is taken to be the population median, the confidence level becomes simply $\Sigma_{i=r+1}^{s-2} \binom{n}{i} 1/2^n$.

It is important to note that the "error" implicit in this method appears only when setting confidence limits for the unknown magnitude of a specified quantile, Q. If the magnitude of Q is hypothesized to be a single specified value, Q', then an exact test of the hypothesis Q = Q' can be made by rejecting if Q' lies outside of the confidence limits $X_r < Q < X_s$.

The methods just discussed establish confidence limits for the unknown magnitude or score below which a *fixed* proportion of the population lies. Binomial methods have also been suggested (3, 6, 31) by which to obtain confidence limits for an unknown population proportion on the basis of the proportion of an obtained sample corresponding to a specified category. These methods, however, appear to be cumbersome, inexact, or both.

g. *Sources.* 19, 32, 40. (See also 3, 6, 9, 21, 22, 31, 33, and I-8 pp. 320-323, 360.)

BIBLIOGRAPHY

1. Arbuthnott, John, An Argument for Divine Providence taken from the constant Regularity observ'd in the Births of both Sexes. Philosophical Transactions of the Royal Society of London, 1710, 27, 186-190.

2. BLYTH, C. R., On efficiencies of the sign test. Technical Report on U. S. Army Ordinance Contract No. DA-11-022-ORD-881, Project No. TB 2-0001 (460).

3. Clopper, C. J. and Pearson, E . S., The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 1934, 26, 404-413.

T 4. Cochran, W. G., The efficiencies of the binomial series tests of significance of a mean and of a correlation coefficient. Journal of the Royal Statistical Society, 1937, 100, 69-73.

* 5. COX, D. R. and STUART, A., Some quick sign tests for trend in location and in dispersion. Biometrika, 1955, 42, 80-95.

6. Crow, E. L., Confidence intervals for a proportion. Biometrika, 1956, 43, 423-435.

7. Dixon, W. J., Power functions of the sign test and power efficiency for normal alternatives. Annals of Mathematical Statistics, 1953, 24, 467-473.

*T 8. DIXON, W. J. and MOOD, A. M., The statistical sign test. Journal of the American Statistical Association, 1946, 41, 557-566.

9. Eisenhart, C., Deming, Lola and Martin, Celia, The probability points of the distribution of the median in random samples from any continuous population. Annals of Mathematical Statistics, 1948, 19, 598-599.

10. Fisher, R. A., *Statistical methods for research workers,* (8th Ed.), London: Oliver and Boyd, 119-120, 1941.

11. Foster, F. G. and Stuart, A., Distribution-free tests in time-series based on the breaking of records. *Journal of the Royal Statistical Society,* 1954, 16, 1-22. See comments by D. R. Cox on page 16.

12. Fraser, D. A. S., Non-parametric theory; scale and location parameters. *Canadian Journal of Mathematics,* 1954, 6, 46-68.

13. Hemelrijk, J., A family of parameterfree tests for symmetry with respect to a given point. I and II. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, (A), 1950, 53, 945-955 and 1186-1198.*

14. Hemelrijk, J., A theorem on the sign test when ties are present. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen,* (A), 1952, 55, 322-326.

15. Hodges, J. L. and Lehmann, E. L., The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics,* 1956, 27, 324-335.

16. Jeeves, T. A. and Richards, R., A note on the power of the sign test. (abstract) *Annals of Mathematical Statistics,* 1950, 21, 618.

* 17. Lehmann, E. L., Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics,* 1951, 22, 165-179.

18. Mosteller, F. and Tukey, J. W., The uses and usefulness of binomial probability paper. *Journal of the American Statistical Association,* 1949, 44, 174-212.

19. Nair, K. R., Table of confidence interval for the median in samples from any continuous population. *Sankhyā,* 1940, 4, 551-558.

T 20. National Bureau of Standards, Tables of the binomial probability distribution, Department of Commerce, National Bureau of Standards, Applied Mathematics Series 6, 1949 (Issued 1950).

21. Noether, G. E., On a connection between confidence and tolerance intervals. Annals of Mathematical Statistics, 1951, 22, 603-604.

22. Noether, G. E., On confidence limits for quantiles. Annals of Mathematical Statistics, 1948, 19, 416-419.

23. Noether, G. E., Sequential tests of randomness, Report No. OSR-TN-54-65, Mathematics division, Boston University, under contract No. AF 18(600)-778, December 1953.

* 24. Noether, G. E., Two sequential tests against trend. Journal of the American Statistical Association, 1956, 51, 440-450.

T 25. Ordinance Corps., Tables of the cumulative binomial probabilities, PB111389, September 1952.

T 26. Princeton University Statistical Research Group, The statistical sign test, (O. S. R. D., 1945, Publ. Bd. No. 23726), Washington, D. C.: US Department of Commerce, 1946, 22 pp. (This is essentially the same as 8)

27. Putter, J., The treatment of ties in some nonparametric tests. Annals of Mathematical Statistics, 1955, 26, 368-386.

T 28. Romig, H. G., 50 - 100 binomial tables, New York: Wiley, 1947.

T 29. Royal Society Mathematical Tables, Vol. III, Table of binomial coefficients, (Ed. by J. C. P. Miller) Cambridge: Cambridge University Press, 1954.

30. Ruist, E., Comparison of tests for non-parametric hypotheses. Arkiv for Matematik, 1954, 3, 133-163.

31. Sandelius, M., A confidence interval for the smallest proportion of a binomial population. Journal of the Royal Statistical Society, (B), 1952, 14, 115-116.

*T 32. Savur, S. R., The use of the median in tests of significance. Proceedings of the Indian Academy of Science, (A), 1937, 5, 564-576.

33. Scheffe, H. and Tukey, J. W., Non-parametric estimation. I. Validation of order statistics. Annals of Mathematical Statistics, 1945, 16, 187-192.

*T 34. STAFF OF THE COMPUTATION LABORATORY, Tables of the cumulative binomial probability distribution, Cambridge, Mass.: Harvard University Press, 1955.

35. Statistical Research Group, Columbia University, Sequential analysis of statistical data: Applications, New York: Columbia University Press, 1945.

36. Stewart, W. M., A note on the power of the sign test. Annals of Mathematical Statistics, 1941, 12, 236-239.

37. Stuart, A., The efficiencies of tests of randomness against normal regression. Journal of the American Statistical Association, 1956, 51, 285-287.

38. Sundrum, R. M., On Lehmann's two-sample test. Annals of Mathematical Statistics, 1954, 25, 139-145.

* 39. Thompson, W. R., On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. Annals of Mathematical Statistics, 1936, 7, 122-128.

40. Wald, A., Sequential method of sampling for deciding between two courses of action. Journal of the American Statistical Association, 1945, 40, 277-306.

41. Wald, A., Sequential tests of statistical hypotheses. Annals of Mathematical Statistics, 1945, 16, 117-186.

42. Walsh, J. E., On the power function of the sign test for slippage of means. Annals of Mathematical Statistics, 1946, 17, 358-362.

T 43. Walsh, J. E., Some bounded significance level properties of the equal-tail sign test. Annals of Mathematical Statistics, 1951, 22, 408-417.

44. Walsh, J. E., Some comments on the efficiency of significance tests. Human Biology, 1949, 21, 205-217.

45. Weinberg, G. H. and Tripp, C. A., A simplification of the sign test. Psychological Bulletin, 1957, 54, 79-80.

T 46. van Wijngarden, A., Table of the cumulative symmetric binomial distribution. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, (A), 1950, 53, 857-868.

See also: Chapter I References 2, 3, 8, 11, 21, 23, 28, 35, 37, 43, 54, 59.

CHAPTER III

THE MULTINOMIAL DISTRIBUTION

The multinomial distribution is important in a study of distribution-free tests because it plays a role in the derivation of a number of exact tests. It is also the exact distribution appropriate to, but too complicated for, the type of test situation in which the chi square statistic is commonly used. Chi square is in fact derived from the multinomial by means of a series of approximations, tantamount to assumptions, which render chi square inexact when sample size is not infinite, and which necessitate considerable skill in applying it properly.

1. Derivation and Assumptions

a. Derivation. Let an event have k possible outcomes, designated by subscripts 1, 2, ..., k, and let these outcomes be mutually exclusive and independent and have probabilities $p_1, p_2, \ldots, p_k$ such that $\sum_{i=1}^{k} p_i = 1$. If the event is allowed to occur n times, the probability that the respective frequencies of occurrence of the various outcomes will be exactly $n_1, n_2, \ldots, n_k$ is

$$\frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \text{ or } n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}.$$

Proof: The probability that the outcomes will occur exactly $n_1, n_2, \ldots, n_k$ times respectively and in a completely specified order (for example, the order in which the first n_1 outcomes are those whose probability is p_1, the next n_2, those whose probability is p_2, etc.) is $p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$. To obtain the probability for these frequencies, but in any order, the preceding product must be multiplied by the number of distinguishable orders. The n outcomes can be permuted in n! ways. But in any one of these permutations, there are n_1 outcomes of the first category which are the same and which can be permuted among themselves in $n_1!$ ways without changing the appearance of the order. And for each of these $n_1!$ permutations, the outcomes of the second category can be permuted with one another in $n_2!$ ways without changing the appearance of the original order, etc. There are thus $n_1!\, n_2! \ldots n_k!$ ways in which each distinguishable order pattern can be permuted without creating a pattern distinguishable from it. Since n! is the number of distinguishable patterns times $n_1!\, n_2! \ldots n_k!$, the number of distinguishable patterns of order is $\frac{n!}{n_1!\, n_2! \ldots n_k!}$ and the probability that in n trials the k categories of outcomes will occur $n_1, n_2, \ldots n_k$ times respectively is

$$\frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

b. Assumptions. Since, in the derivation , the same value, p_i was taken as the probability for outcome i in each of its n_i occurrences,

p_i must not vary from event to event. The outcome of a given event must therefore be independent of the outcomes of any of the n-1 other events. Not only must the probabilities of the various possible outcomes of an immediate event be independent of the actual, observed, outcomes of the previous events, they must also be mutually exclusive. This assumption is necessitated by the fact that the probability of a given set of n outcomes was obtained in the derivation by taking the product of the n individual outcome probabilities; to obtain compound probability in this fashion, the individual probabilities must be mutually exclusive. (See Mood I, 30-36). Another assumption is that

$\sum_{i=1}^{k} n_i = n.$ Unless this is the case, $\frac{n!}{n_1!\, n_2! \,\ldots\, n_k!}$ does not give the number of distinguishable orders of obtained outcomes as required in the derivation, and, in fact becomes meaningless. Since k mutually exclusive outcomes are recognized as possible,

$\sum_{i=1}^{k} p_i$ must equal 1. Otherwise a real probability, $1 - \sum_{i=1}^{k} p_i$, would exist for outcomes in an additional category or categories not considered. (Furthermore, the occurrence of such uncategorized outcomes would mean that n would be greater than $\sum_{i=1}^{k} n_i$.) Finally, since p_1, p_2, etc. are chance probabilities, sampling must be random, i.e., the n events or trials must be selected on a chance basis from the infinite number of potential events available. Specifically this means, among other things, that no bias shall have operated to exclude valid but "unfavorable" data from the test.

Use of the multinomial distribution in statistical tests requires that the probabilities for all of the possible outcomes be known exactly and be included in the formula $n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}$. It is important, however, to recognize that the experimenter is free to define both the sample space in which he is interested and the categories which divide that sample space into k mutually exclusive parts. The experimenter must, in fact, be careful to do this in such a way as to define precisely that situation in which he is interested. If he fails he will obtain an exact probability for a situation in which he is not interested, and this probability will differ, perhaps considerably, from the exact probability for the situation in which his interest lies. For example, in coin tossing, in addition to "heads" and "tails", the outcome category "on the rim" has finite probability which usually cannot be specified. Therefore, although heads and tails have equal probabilities, these probabilities are unknown since their sum is not 1. By defining his sample space as that including only those outcome categories in which the coin lands flat, the experimenter enables himself to specify as .50 the probability of heads and the probability of tails. The experimenter is no better off, however, unless his interest is confined to the sample space consisting only of heads and tails, i. e., is confined to the frequency of heads relative to tails rather than tosses. Again, the experimenter may be interested in broader categories than those into which his data are fitted. In such cases he should use the categories in which he is interested rather than those in which the data are available. For example, in tossing two coins simultaneously the possible outcomes will be defined to be two heads ($P_r = 1/4$), a head and a tail ($P_r = 1/2$), and two tails ($P_r = 1/4$). Suppose that the two coins have been simultaneously tossed n times and that the frequencies of the respective outcomes named above are n_1, n_2 and n_3. If the experimenter is interested in the point probability of the obtained frequencies for the outcomes stated, the proper formula is $\frac{n!}{n_1!\, n_2!\, n_3!} (1/4)^{n_1+n_3} (1/2)^{n_2}$. On the other hand, if he is interested in the probability of the obtained frequencies for the recategorized outcomes, "coins have same side up" ($P_r = 1/2$) and "coins have different sides up" ($P_r = 1/2$), the obtained frequencies are n_1+n_3 and n_2 respectively and the probability is $\frac{n!}{(n_1+n_3)!\, n_2!} (1/2)^{n_1+n_3} (1/2)^{n_2}$. The probabilities for the same data under the two different categorizations of outcome are not the same:

$$\frac{n!}{n_1!\, n_2!\, n_3!} (1/4)^{n_1+n_3} (1/2)^{n_2} \overset{?}{=} \frac{n!}{(n_1+n_3)!\, n_2!} (1/2)^{n_1+n_3} (1/2)^{n_2}$$

$$\frac{1}{n_1!\, n_3!} (1/4)^{n_1+n_3} \overset{?}{=} \frac{1}{(n_1+n_3)!} (1/2)^{n_1+n_3}$$

$$\frac{1}{n_1!\, n_3!} (1/2)^{n_1+n_3} \overset{?}{=} \frac{1}{(n_1+n_3)!}$$

$$\frac{(n_1+n_3)!}{n_1!\,n_3!} \overset{?}{=} 2^{n_1+n_3}$$

$$\binom{n_1+n_3}{n_1} \overset{?}{=} 2^{n_1+n_3}$$

Substituting N for n_1+n_3, the questioned equality becomes $\binom{N}{n_1} \overset{?}{=} 2^N$ which is obviously absurd since $\binom{N}{n_1}$ varies with the particular values of n_1 and n_3, while 2^N does not, varying only with their sum. The reason for the discrepancy between the two probabilities is that one states merely that n_1+n_3 tosses result in either two heads or two tails without specifying precisely how many of these shall be two heads; the other probability *does* specify this further and much more restrictive information. The latter probability is, therefore, much smaller than the former.

The multinomial distribution is seldom used directly as the basis of a statistical test. This is partly attributable to the fact that the exact probabilities for the various outcome categories, although required by the test, are seldom known; and it is partly because, unless n is quite small, computation of cumulative probabilities, i.e., significance levels, is likely to be extremely time consuming. Nor is this distribution extensively tabled except for the special case where $k=2$, i.e., except for the case of the binomial distribution. The reason for the lack of extensive tables is obvious: the number, $2k - 1$, of required parameters is prohibitively large.

2. The Chi Square Approximation to the Multinomial

Because chi-square occupies a prominent position in most elementary statistical texts it will be assumed that the details of its application are familiar to the reader. Because it is one of the most misunderstood and misused of statistical tests, its theory and the hazards of its *mis*application will be discussed in detail.

The chi-square distribution is derived from the multinomial, three approximations being required in the derivation and therefore qualifying the use of chi-square. The first approximation consists in replacing the factorials in the multinomial

$$\frac{n!}{n_1!\, n_2! \ldots n_k!}\; p_1^{n_1}\, p_2^{n_2} \ldots p_k^{n_k}$$ by their Stirling approximations.

The second approximation "is similar in character to the familiar one by which an expression of the form $(1+x/m)^m$ is replaced by e^x when m is large" (27). The final approximation consists in replacing by an integral the discrete summation representing the cumulative distribution function.

Each of these three approximations presupposes infinite, i.e. very large, n's and becomes increasingly poor with diminishing sample size. Each is strictly valid only for samples of infinite size.

The first two approximations together are equivalent to substituting for the multinomial distribution its multivariate normal approximation. At this point the assumption is necessitated that, for each category, the observed frequencies are normally distributed about the expected frequency as a mean. For a single multinomial category, outcomes are binomially distributed; therefore replacing the multinomial distribution by its multivariate normal approximation is equivalent to substituting the univariate normal distribution for the true binomial distribution of outcomes within each multinomial category. In fact, the working formula by which data are referred to the chi square tables can, for the case of one degree of freedom, be easily derived by making this substitution. Consider a binomial variate with the probability p for a single event. The point probability that it will occur r times in n trials is $\frac{n!}{r!\,(n-r)!}\; p^r\,(1-p)^{n-r}$, or, if the normal approximation is used, the corresponding cumulative probability is that of the "normal" deviate $\chi = \frac{r-np}{\sqrt{np(1-p)}}$, np being the mean and $\sqrt{np(1-p)}$ the standard deviation of the binomial distribution. If both sides of the equation are squared and numerator and denominator of the right side are multiplied by n, it becomes

$\chi^2 = \frac{n(r-np)^2}{np(n-np)}$. Now substitute f_{o_1} for r and f_{e_1} for np, giving

$$\chi^2 = \frac{n(f_{o_1} - f_{e_1})^2}{f_{e_1}(n-f_{e_1})} = \frac{(f_{o_1} - f_{e_1})^2}{f_{e_1}} + \frac{(f_{o_1} - f_{e_1})^2}{n-f_{e_1}}$$

$$= \frac{(f_{o_1} - f_{e_1})^2}{f_{e_1}} + \frac{(n-f_{o_1} - n+f_{e_1})^2}{n-f_{e_1}} .$$

If now f_{e_2} is substituted for $n-f_{e_1}$ and f_{o_2} for $n-f_{o_1}$,

$$\chi^2 = \frac{(f_{o_1} - f_{e_1})^2}{f_{e_1}} + \frac{(f_{o_2} - f_{e_2})^2}{f_{e_2}}$$ which is the formula used to calculate χ^2 with one degree of freedom from data in which f_{o_1} and f_{e_1} are the observed and expected frequencies of occurrence and f_{o_2} and f_{e_2} are the corresponding frequencies of nonoccurrence. (It is easily seen from the foregoing that chi is normally distributed when chi square is based on a single degree of freedom.)

The assumption that observed frequencies are normally distributed about their expected frequency is, of course, incapable of being met exactly unless n is infinite at which point the binomial distribution and its asymptotic normal "approximation" are identical. The normality assumption is therefore equivalent to the "assumption" that n is infinite, or, since the expected frequency, f_{e_i}, equals np_i, that all expected frequencies are infinite.

In more practical terms, the "assumption" of normal distribution of observed frequencies will be negligibly violated if the following conditions exist: (a) n is so large that for every $p_i \neq .50$, the true, i.e. binomial, distribution of observed frequencies within each category has no more than negligible asymmetry; this must be the case if the binomial is to be well approximated by the "fitted" normal distribution which is symmetrical, (b) n is so large that for each category the area of the "fitted" normal curve covering impossible "observed" frequencies, i.e., those frequencies which are less than zero or greater than n, is negligible relative to the size α of the significance level being used for the chi square test, (c) n is so large that if for each category the points corresponding to observed frequencies in the binomial distribution of observed frequencies were connected by line segments, the result would have the appearance of a smoothly continuous curve. The smaller the smallest $p_i$is, the larger n must be to produce the effects named; and the smaller the significance level chosen for the chi square test, the greater the relative importance of asymmetry, the alleged probability of impossible frequencies, and discontinuity, and therefore the larger n must be to make these effects negligible. The term "negligible" has not been, and will not be defined. Any subjective definition will suffice if consistently applied, since, in the above discussion, that degree of cause which is defined as negligible will have an effect whose degree is of about the same order of negligibility.

Much acrimonious controversy has raged over the question of how small an expected frequency can be safely used in a chi-square test. The reason for the animosity is not hard to find. Since for any expected frequency short of infinity, chi square is an approximation rather than an exact test, the question of how small an expected frequency can be tolerated resolves itself into a pure matter of opinion as to how close an approximation is "good". And most writers have not quantified the degree of approximation which they find tolerable other than by specifying a minimum acceptable expected frequency. The most popular rule of thumb appears to be that "no expected frequency should be less than 5", possibly because the normal approximation to the binomial is regarded as good if np exceeds 5. However, such rules overlook the fact that the effect of an assumption violation is usually a function of several factors only one of which, i. e expected frequency, is mentioned in the rule. For example, there is every reason to believe a priori that (a) the variance and degree of symmetry of the sampling distribution of observed frequencies, (b) the "height" of the significance level chosen, and (c) the number of categories, will be important factors in determining whether or not the use of an expected frequency as low as 5 will have an appreciable effect upon the closeness of approximation of the chi square significance level to the "true" multinomial significance level. The smaller the variance of the true sampling distribution of observed frequencies the smaller will be the area of the normal distribution, assumed for them, which occupies the region corresponding to negative, and therefore impossible, frequencies. And the more nearly symmetrical the sampling distribution of observed frequencies (i. e., the closer p is to .50 for a given n), the better it will be approximated by the normal distribution it is "assumed" to have. Curve "fits" are usually poorest at their tails, therefore the distortion of the chi-square approximation should be greater the higher the significance level. Finally, since chi-square is the sum of squared deviations divided by the respective expected frequencies, the effect of a single very small expected frequency in a large number of categories would exert a smaller relative influence upon the sum, and therefore chi-square, than would be the case if a smaller number of categories were being used. Tables III and IV show the distorting effect of some of these factors upon chi square probabilities when the expected frequency is 5 and 2 respectively. For other studies of the sensitivity of chi square to gross violations of its assumptions, see (9, 36, 56, 59, 66).

The prohibition against small expected frequencies has led to the widely accepted practice of pooling categories in order to bring the expected frequencies for the combined categories up to the required size. Such pooling, however, involves an arbitrary decision which must usually be made subsequent to the collection of data. Such a posteriori manipulation of test parameters, i.e. categories, in effect violates the assumption of random sampling since outcomes are being influenced by a factor other than chance. This objection is not an academic one, since the manner in which categories are combined can dramatically affect the significance levels obtained for a given set of data. Gumbel (29) gives an example of a goodness of fit test in which probability levels calculated by chi-square from the same data, using the same abscissa interval length to define categories (and of course the same number of categories in each case), vary by a factor of 30 depending on the point chosen for the beginning of the first interval. When dealing with contingency tables the expected frequencies are usually not known in advance of sampling, being calculated from the marginal observed frequencies. In such cases the experimenter may be forced to choose between a posteriori pooling and using too small an expected frequency, assumptions being violated under either alternative. However, in testing goodness of fit to a completely known and tabled continuous function the issue can be avoided because sufficient information is available to set, in advance, the minimum expected frequency which the experimenter is willing to tolerate. The "X-axis" of the distribution to which fit is being tested is divided into k intervals so that the area under the curve above each interval is the same for every interval, each such area therefore equaling 1/k. Each interval therefore is a category whose probability is 1/k, and if n observations are taken, the expected frequency for each category is n/k. (See 42 and 63)

TABLE III

Influence of Sample Size, n, Category Probability, p, and Approximate Significance Level, α, upon the Error in Chi Square Probabilities when Expected Frequency, $f_e = np$, is 5 and there Are Only Two Observed Frequencies, i.e., Chi Square Has One Degree of Freedom

n	p	α	f_o [Smallest $f_o > np$ for which true cumulative $Pr(\lvert f_o - f_e \rvert) \leq \alpha$]	Cumulative $Pr(\lvert f_o - f_e \rvert)$ true	χ^2	χ^2 with Yates' correction χ_y^2	$\frac{\text{true}}{\chi^2}$	$\frac{\text{true}}{\chi_y^2}$
10	.50	.05	9	.02148	.01141	.02683	1.88	.80
		.01	10	.00196	.00157	.00443	1.25	.44
		.001						
20	.25	.05	10	.01703	.00982	.02013	1.73	.85
		.01	11	.00394	.00195	.00451	2.02	.87
		.001	12	.00094	.00030	.00079	3.13	1.19
50	.10	.05	10	.02969	.01842	.03392	1.61	.88
		.01	11	.00935	.00468	.00951	2.00	.98
		.001	13	.00100	.00016	.00041	6.25	2.44
100	.05	.05	10	.03411	.02179	.03892	1.57	.88
		.01	12	.00427	.00132	.00286	3.23	1.49
		.001	14	.00046	.00004	.00010	11.50	4.60
500	.01	.05	10	.03767	.02464	.04307	1.53	.87
		.01	12	.00521	.00166	.00348	3.14	1.50
		.001	14	.00065	.00005	.00013	13.00	5.00

True probabilities were obtained from binomial tables with cumulative $Pr(\lvert f_o - f_e \rvert) = Pr(r \geq np + \lvert f_o - np \rvert) + Pr(r \leq np - \lvert f_o - np \rvert)$. Chi square cumulative $Pr(\lvert f_o - f_e \rvert)$ was obtained from normal tables since, for 1 degree of freedom, $\chi = \frac{f_o - np}{\sqrt{np(1-p)}}$ is a normal deviate (with Yates' correction $\chi = \frac{\lvert f_o - np \rvert - 1/2}{\sqrt{np(1-p)}}$).

TABLE IV

Influence of Sample Size, n, Category Probability, p, and Approximate Significance Level, α, upon the Error in Chi Square Probabilities when Expected Frequency, $f_e = np$, is 2 and there Are Only Two Observed Frequencies, i.e., Chi Square Has One Degree of Freedom

n	p	α	f_o [Smallest $f_o > np$ for which true cumulative $Pr(\lvert f_o - f_e\rvert) \leq \alpha$]	Cumulative $Pr(\lvert f_o - f_e\rvert)$				
				true	χ^2	χ^2 with Yates' correction χ^2_y	$\frac{\text{true}}{\chi^2}$	$\frac{\text{true}}{\chi^2_y}$
10	.20	.05	5	.03279	.01769	.04815	1.85	.68
		.01	6	.00637	.00157	.00566	4.06	1.13
		.001	7	.00086	.00008	.00037	10.75	2.32
20	.10	.05	5	.04317	.02535	.06246	1.70	.69
		.01	7	.00239	.00019	.00080	12.58	2.99
		.001	8	.00042	.0000077	.00004	54.54	10.50
50	.04	.05	5	.04897	.03039	.07123	1.61	.69
		.01	7	.00361	.00031	.00116	11.65	3.11
		.001	8	.00078	.000015	.00007	52.00	11.14
100	.02	.05	6	.01548	.00428	.01242	3.62	1.25
		.01	7	.00406	.00036	.00131	11.28	3.10
		.001	8	.00093	.00002	.00009	46.50	10.33
200	.01	.05	6	.01602	.00447	.01288	3.58	1.24
		.01	7	.00430	.00038	.00138	11.32	3.12
		.001	9	.00021	.0000007	.0000039	300.00	53.85

The last of the three approximations used in the derivation of the chi square density function consisted of replacing a discrete sum by an integral. The result is that the tabled chi square distribution is continuous while the multinomial distribution which it approximates is discretely distributed as is the "working formula",

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} ,$$

by which "chi square" is calculated from obtained data. Substituting an integral for a discrete summation is conscionable only when the discrete function involves so many discrete values, each differing so slightly from the adjacent values, as to be well approximated by a continuous function. When expected frequencies are small, the number of different values which the observed frequencies can assume is quite limited, and this discrete distribution is not well approximated by the continuous chi-square distribution. However, when chi square is based upon a single degree of freedom, the approximation can generally be improved by applying Yates' correction for continuity (66). This consists of reducing the absolute value of the deviations of observed from expected frequencies by 1/2 prior to squaring them in the calculation of chi square. The correction does not compensate exactly for the discontinuity in the sampling distribution from which the obtained data were "drawn"; it may, in fact, aggravate rather than reduce the error. "In symmetrical and nearly symmetrical distributions" ... the correction overestimates the true... "probabilities at both tails and under-estimates them near the centre of the distribution. Such discrepancies, however, are small compared with those arising in violently unsymmetrical cases." (66) Generally Yates' correction is an improvement. It is commonly recommended for calculating chi squares based on one degree of freedom, (except when $/f_o - f_e/ \leq 1/2$, in which case it "overcorrects"). It should not be used, however, in calculating individual chi squares, with one degree of freedom, which are to be added, and their degrees of freedom summed, to obtain a total chi-square. (See Chapter IV for a superior method in the case of certain fourfold contingency tables.)

Since the multinomial distribution from which chi-square is derived applies only to repeated independent events the chi-square test is equally dependent upon the assumption that *each* of the occurrences of an event comprising a frequency of occurrence is *independent* of all other occurrences of the event. This is one of the most frequently violated assumptions of chi square (40). Also traceable to the multinomial is the assumption that *outcome categories are mutually exclusive.*

If p_i is the probability that a single event will have an outcome which places it in the i^{th} category, then the expected frequency, f_{e_i}, for that category is np_i, where n is the number of times the event is permitted to occur. Since $p_i = f_{e_i}/n$ and since $\Sigma p_i = 1$, it follows that $\Sigma f_{e_i} = n$. Obviously if f_{o_i} is the observed frequency for category i, then $\Sigma f_{o_i} = n$, or $\Sigma f_{o_i} = \Sigma f_{e_i}$. This is frequently stated as an assumption: *the sum of the observed frequencies must equal the sum of the expected frequencies.* It is probably most frequently violated by failing to give the f_{e_i} the exact decimal values calculated for them, rounding them off instead to whole numbers.

Another assumption is that the introduction of information concerning higher moments, such as the variance of the distribution in a test of fit, does not alter the condition expressed by the equality $\Sigma(f_{o_i} - np_i) = 0$. This is expressed in the requirement that necessary equations involving the above equality are *linear and homogeneous* in the variables $(f_{o_i} - np_i)$. (27)

When useful information can be introduced into the chi-square test, such as the variance of a distribution whose "goodness of fit" is being tested, the effect is to identify and specify the particular values which the chi variate may assume in one of the dimensions of the hyper-space which the chi distribution occupies. The effect of each such "restriction" is to reduce by one the number of dimensions in which chi is "free" to vary. The number of such free dimensions is known as the number of degrees of freedom. Fisher (23) presents the rationale for this reduction as follows. "The common sense of this correction lies in the fact that when the population with which the sample is compared has been artificially identified with the sample in certain respects, such as marginal frequencies, or the moments, we shall evidently make an exaggerated estimate of the closeness of agreement between sample and population, if we regard the sample as an unselected sample of a population known *a priori.*"

Chi square, although deceptively simple in application, is one of the most complicated statistics in its theoretical basis. It has been widely misunderstood by professional statisticians as well as by laymen. Nearly a quarter of a century elapsed after Pearson's publication of the original article on chi square before statisticians understood how degrees of freedom are affected by linear restrictions upon the data. And in a survey (40) of the use of chi square by psychologists publishing in a professional journal, in nine out of fourteen articles the application of chi square was found to be "clearly unwarranted". As a symptom of the confusion surrounding its use, extended discussion and debate has surrounded such questions as the correct number of degrees of freedom (8, 20, 22, 23, 24, 39, 40, 48, 67) the minimum tolerable expected frequency (8, 19, 39, 40) when and how to apply Yates' correction (1, 8, 11, 40, 66), and even whether or not the hypothesis of "fit" should be rejected when the fit is so good as to be expected rarely (2, 6, 8, 58). (Curiously enough the affirmative in the last named controversy was taken by no less a statistician than R. A. Fisher; it is effectively and eloquently rebutted by Stuart (58)).

Aside from its complexity chi-square suffers from a number of practical and theoretical shortcomings. Whether or not an hypothesis of fit will be rejected may depend as much upon the statistician as upon the obtained data, since probabilities may be greatly affected, a posteriori, by the manner in which the data are grouped into "intervals" or cells. Since all deviations are squared in the computation of chi-square, the test is completely insensitive to the directions of the deviations, regarding a series of unidirectional deviations as no more significant than a set of deviations, varying haphazardly in direction from the hypothesized curve but having the same absolute magnitudes. Applications of chi-square in which, for a given sample size, all expected frequencies can be specified in advance of sampling are relatively rare. However, it is only in such cases that the chi-square test is truly parameter-free. In all other cases chi square is parametric in the sense that population parameters, e.g., expected frequencies in a contingency table or the variance of a "fitted" distribution, must be estimated a posteriori from sample data. And in such applications the excellence of the test, i.e., the accuracy of its calculated probabilities, depends upon the efficiency of the estimates and upon their accuracy in the particular case in question. In the sense that chi-square assumes the sampling distribution of observed frequencies in each category to be normally distributed, it is not "distribution-free". More accurately phrased, chi square falsely assumes a multivariate normal distribution in cases where the true distribution must necessarily be the multinomial. Because of its resort to such approximations, it is an inexact test.

Because of its many shortcomings, other tests, such as the

Kolmogorov-Smirnov test of fit will, in most cases, be preferable. In some few cases chi square may be desirable because of its additive property or because of its ability to make allowance for the identification of parameters in the hypothesized population on the basis of data whose fit is being tested. However, unless such unique properties are required, it will be wise to seek another test; and, when its avoidance is impossible, chi square should be used with great caution.

SUMMARY

The multinomial test assumes random sampling of events whose outcomes are independent and fall into mutually exclusive categories, the sum of whose probabilities is unity. The test yields probabilities which, for a given set of data, vary with the system of categorization used. The practical validity of the test therefore depends upon defining and establishing categories which correspond precisely with the situation to which the experimenter wishes to extend statistical inference. Although it is an exact test, it may require prohibitively extensive computation, especially when n is large, since tables are not available for the case of more than two categories.

The chi square test is extensively tabled and was designed for those situations in which the multinomial test would be appropriate if computation of probabilities were easier. The chi square distribution was, in fact, derived from the multinomial distribution, the derivation having entailed three asymptotically valid approximations. It is the asymptotic distribution for the statistic $\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$, which, at finite sample sizes is an inexact statistic.

Because of its relationship to the multinomial, the chi square test incorporates all of the assumptions on which the multinomial is based. It therefore assumes that events are randomly sampled, that possible outcomes, i.e. categories, are mutually exclusive, that actual outcomes are independent, that $\Sigma p_i = 1$ and $\Sigma f_o = \Sigma f_e$.

Further assumptions are required due to steps taken in the derivation of chi square. Within each multinomial category the frequency of occurrence is binomially distributed with mean equal to np. However, the derivation regards this frequency as normally distributed. The chi square test therefore makes the assumption that within each chi square "cell" the population of "observed" frequencies is normally distributed about the expected frequency, np, as a mean. This is equivalent to assuming infinite n, since it is only for that case that the binomial can be exactly fitted by a normal distribution, and since $f_e = np$, it is equivalent to assuming infinite expected frequencies. Another assumption, traceable to the derivation is that all restrictions on the data are both linear and homogeneous.

Chi square will not be a good approximate test unless the binomial distribution of observed frequencies within each category is well approximated by a normal distribution. The normal approximation worsens with increasingly remote tail positions, with increasing asymmetry of the binomial and with decreasing sample size. Therefore the accuracy of the chi square test is a function of α, the significance level, p_i, the probability that a single event will have an outcome in the i^{th} category, and n, the total number of events. The rule that no expected frequency, $f_e = np$, should be less than 5 is a poor one since the accuracy of the chi square test varies widely with the individual values of n and p as well as with their product and since the rule says nothing about α.

The tabled chi square distribution is a continuous one. The distribution of the value, $X^2 = \Sigma \dfrac{(f_o - f_e)^2}{f_e}$, by which "chi square" is calculated from obtained data, must, however, have a discrete distribution since observed frequencies are necessarily integers. This introduces an error which can usually be reduced, but is not entirely removed, by applying Yates' correction for continuity when chi square is based upon a single degree of freedom. It should not be applied, however, if the individual chi squares are to be added to obtain a total chi square.

If "natural" categories are combined or "pooled" in order to increase the size of expected frequencies or in order to shorten computations, the redefinition of categories changes the situation to which "fit" is being tested. It therefore alters the null hypothesis in a way which is fairly obvious in the case of contingency tables, more subtle in the case of tests for goodness of fit (where the null hypothesis actually being tested is that the various categories have the expected frequencies assigned to them, not that the two "curves" are identical). This combining of categories may obscure a real effect and lead to "acceptance" when the uncombined data actually call for "rejection" of the hypothesis in which the experimenter actually is interested, or it may do the opposite. Furthermore, in tests of goodness of fit to a continuous distribution, not only will the choice of interval length affect the obtained significance level, but even the choice of the point at which to begin the leftmost or rightmost abscissa interval may have a profound effect upon the significance level obtained. In fact profound effects may attend <u>any</u> situation in which categories are determined on the basis of a posteriori expediency rather than by a "natural" discrimination between precisely those event outcomes in which the experimenter is interested.

Although chi square is extremely complicated in its derivation, its simplicity of actual computational application has made it a favorite among the statistically naive. This treacherous combination of theoretical complexity and deceptive simplicity in practical application has made it a perennially misused statistic. Even mathematical statisticians, including those originating it and modifying it, have experienced great difficulty in determining its proper use and even greater lack of success in explaining it to lay statisticians. Therefore research workers will be well advised to check thoroughly into the theoretical admissibility of any contemplated application of this statistic. Those not possessing the requisite sophistication for such an undertaking are urged to <u>shun chi square.</u>

BIBLIOGRAPHY

1. ADLER, F., Yates' correction and the statisticians. Journal of the American Statistical Association, 1951, 46, 490-501.

2. Berkson, J., Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 1938, 33, 526-536.

T 3. Bliss, C. I., A chart of the chi-square distribution. Journal of the American Statistical Association, 1944, 39, 246-248.

4. Borsting, J., On the addition of chi-squares. (Abstract) Annals of Mathematical Statistics, 1952, 23, 480.

5. Brownlee, J., Some experiments to test the theory of goodness of fit. Journal of the Royal Statistical Society, 1924, 87, 76-82.

6. Camp, B. H., Further interpretations of the chi-square test. Journal of the American Statistical Association, 1938, 33, 537-542.

7. Chernoff, H. and Lehmann, E. L., The use of maximum likelihood estimates in χ^2 tests for goodness of fit. Annals of Mathematical Statistics, 1954, 25, 579-586.

8. COCHRAN, W. G., The χ^2 test of goodness of fit. Annals of Mathematical Statistics, 1952, 23, 315-345.

9. Cochran, W. G., The statistical analysis of field counts of diseased plants. Journal of the Royal Statistical Society, 1936, , 49-67.

10. Craig, C. C., Combination of neighboring cells in contingency tables. Journal of the American Statistical Association, 1953, 48, 104-112.

11. Crow, E. L., Some cases in which Yates' correction should not be applied. Journal of the American Statistical Association, 1952, 47, 303-304.

* 12. David, F. N., A X^2 'smooth' test for goodness of fit. Biometrika, 1947, 34, 299-310.

13. David, F. N., An alternative form of X^2. Biometrika, 1950, 37, 448-451.

14. David, F. N., Correlations between X^2 cells. Biometrika, 1948, 35, 418-422.

15. David, F. N., On Neyman's smooth test for goodness of fit I. Distribution of the criterion X^2 when the hypothesis is true. Biometrika, 1939, 31, 191-199.

16. Dawson, R. B., A simplified expression for the variance of the X^2- function on a contingency table. Biometrika, 1954, 41, 280.

17. Deming, W. E., Some thoughts on curve fitting and the chi test. Journal of the American Statistical Association, 1938, 33, 543-551.

18. Deming, W. E., The chi-test and curve fitting. Journal of the American Statistical Association, 1934, 29, 372-382.

19. Edwards, A., On "The use and misuse of the chi-square test" - the case of the 2 X 2 contingency table. Psychological Bulletin, 1950, 47, 341-346.

20. Fisher, R. A., Bayes' theorem and the fourfold table. Eugenics Review, 1926, 18, 32-33.

21. Fisher, R. A., On a property connecting the X^2 measure of discrepancy with the method of maximum likelihood. Atti del Congresso Internazionale dei Mathematici, Bologna, 1928, 6, 95-100.

22. FISHER, R. A., On the interpretation of chi-square from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, 1922, 85, 87-94.

23. Fisher, R. A., Statistical tests of agreement between observation and hypothesis. Economica, 1923, 8, 1-9.

24. Fisher, R. A., The conditions under which X^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society,* 1924, 87, 442-450.

25. Fraser, D. A. S., Note on the X^2 smooth test. *Biometrika,* 1950, 37, 447-448.

26. Freeman, G. H. and Halton, J. H., Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika,* 1951, 38, 141-149.

27. FRY, T. C., The X^2 test of significance. *Journal of the American Statistical Association,* 1938, 33, 513-525.

28. GREENHOOD, E. R., *Detailed proof of the chi-square test of goodness of fit,* Cambridge Mass.: Harvard University Press, 1940.

29. Gumbel, E. J., On the reliability of the classical chi-square test. *Annals of Mathematical Statistics,* 1943, 14, 253-263.

30. Haldane, J. B. S., The exact value of the moments of the distribution of X^2 used as a test of goodness of fit, when expectations are small. *Biometrika,* 1937, 29, 133-143.

31. Hoel, P. G., On the chi-square distribution for small samples. *Annals of Mathematical Statistics,* 1938, 9, 158-165.

32. Irwin, J. O., Note on the subdivision of X^2 into components. *Biometrika,* 1949, 36, 130-134.

33. Kimball, A. W., Short-cut formula for the exact partition of χ^2 in contingency tables. *Biometrics,* 1954, 10, 452-458.

34. Lancaster, H. O., A reconciliation of X^2 considered from metrical and enumerative aspects. *Sankhyā,* 1953, 13, 1-10.

35. Lancaster, H. O., Complex contingency tables treated by the partition of X^2. *Journal of the Royal Statistical Society,* (B), 1951, 13, 242-249.

36. Lancaster, H. O., Statistical control of counting experiments. *Biometrika,* 1952, 39, 419-422.

37. Lancaster, H. O., The derivation and partition of X^2 in certain discrete distributions. Biometrika, 1949, 36, 117-129.

38. Lancaster, H. O., The exact partition of X^2 and its application to the problem of the pooling of small expectations. Biometrika, 1950, 37, 267-270.

39. Lewis, D. and Burke, C. J., Further discussion of the use and misuse of the chi square test. Psychological Bulletin, 1950, 47, 347-355.

40. LEWIS, D. and BURKE, C. J., The use and misuse of the chi-square test. Psychological Bulletin, 1949, 46, 433-489.

T 41. Lewis, T., 99.9 and 0.1% points of the X^2 distribution, Biometrika, 1953, 40, 421-426.

42. Mann, H. B. and Wald, A., On the choice of the number of class intervals in the application of the chi square test. Annals of Mathematical Statistics, 1942, 13, 306-317.

43. Neymann, J., Contributions to the theory of the X^2 test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles: University of California Press, 1949, pp. 239-273.

44. Neyman, J., Empirical comparison of the "smooth" test for goodness of fit with the Pearson's X^2 test. (Abstract) Annals of Mathematical Statistics, 1940, 11, 478.

* 45. Neyman, J., "Smooth test" for goodness of fit. Skandinavisk Aktuarietidskrift, 1937, , 149-199.

46. Neyman, J. and Pearson, E. S., on the use and interpretation of certain test criteria for purposes of statistical inference Part II Biometrika, 1928, 20A, 263-294.

47. Pastore, N., Some comments on "The use and misuse of the chi-square test." Psychological Bulletin, 1950, 47, 338-340.

48. Pearson, K., On the X^2 test of goodness of fit. Biometrika, 1922, 14, 186-191.

* 49. Pearson, K., On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. Philosophical Magazine, 5 series, 1900, 50, 157-175.

50. Pearson, K., On the general theory of multiple contingency with special reference to partial contingency. Biometrika, 1916, 145-158.

51. Pearson, K., On the probability that two independent distributions of frequency are really samples from the same parent population. Biometrika, 1932, 24, 457-470.

52. Pearson, K., On the probability that two independent distributions of frequency are really samples from the same populations. Biometrika, 1911, 8, 250-254.

53. Peters, C. C., The misuse of chi-square - a reply to Lewis and Burke. Psychological Bulletin, 1950, 47, 331-337.

54. Robinson, S., An experiment regarding the χ^2 test. Annals of Mathematical Statistics, 1933, 4, 285-287.

55. Seal, H. L., A note on the χ^2 smooth test. Biometrika, 1948, 35, 202.

56. El Shanawany, M. R., An illustration of the accuracy of the χ^2 approximation. Biometrika, 1936, 28, 179-187.

57. Shewhart, W. A., Economic control of quality of manufactured product, New York: van Nostrand, 1931, pp. 128-138.

58. Stuart, A., Too good to be true? Applied Statistics, 1954, 3, 29-32.

59. Sukhatme, P. V., On the distribution of χ^2 in samples of the Poisson series. Journal of the Royal Statistical Society, (B), 1938, 5, 75-79.

T 60. Thompson, Catherine M., Table of percentage points of the χ^2 distribution. Biometrika, 1941, 32, 187-191.

61. Walker, Helen M., Degrees of freedom. Journal of Educational Psychology, 1940, 31, 253-269.

62. Wilks, S. S., The likelihood test of independence in contingency tables. Annals of Mathematical Statistics, 1935, 6, 190-196.

63. WILLIAMS, C. A., On the choice of the number and width of classes for the chi-square test of goodness of fit. Journal of the American Statistical Association, 1950, 45, 77-86.

T 64. WILSON, E. B., An introduction to scientific research, New York: McGraw-Hill, 1952, pp. 197-202, 229-231.

65. Wishart, J., χ^2 probabilities for large numbers of degrees of freedom. Biometrika, 1956, 43, 92-95.

66. YATES, F., Contingency tables involving small numbers and the χ^2 test. Journal of the Royal Statistical Society, (B), 1934, 1, 217-235.

67. Yule, G. U., On the application of the χ^2 method to association and contingency tables, with experimental illustrations. Journal of the Royal Statistical Society, 1922, 85, 95-104.

CHAPTER IV

EXACT TREATMENT OF FREQUENCY DATA IN FOURFOLD TABLES

A test statistic having a "binomial" derivation (but not a binomial distribution) can be used to test whether or not two samples dichotomized into A's and B's came from populations with equal A/B ratios. Tests of this type use only frequency data and are easy to apply. Depending upon the choice of dichotomous categories, the method may be used to test for equal A/B ratios, or may be used to test for location, dispersion, correlation, or trend. The method can be regarded as an application of Fisher's Method of Randomization (See next chapter) to observation frequencies rather than their magnitudes; and, in this context, it is of historical importance in the development of distribution-free methods.

1. Fisher's Exact Method

a. Rationale. Suppose that two populations, differing perhaps in many ways, nevertheless each consist entirely of units which belong to one or the other of two mutually exclusive categories, A and B. Suppose further that a sample has been drawn from each population and the experimenter wishes to test the hypothesis that the proportion of A's in Population I is the same as that in Population II. Letting the frequency data be represented by the table shown below,

	Category		
	A	B	Total
Sample I	a	b	m
Sample II	c	d	n
Total	r	s	N

if the hypothesis is true one would expect cell frequency a to be such that, on the average, the proportion, a/m of A's in Sample I would equal the proportion, c/n, of A's in Sample II. Therefore one might reasonably reject the null hypothesis of equal proportions of A's, at the α level of significance, if the obtained cell frequency a is among that proportion, α, of possible values of a which cause a/m to differ from c/n by the greatest amount.

If the validity of the hypothesis be accepted, it follows that the true proportion of A's among the A's and B's in Population I, in Population II and in both populations combined, is the same. Let p be this common, but unknown, proportion. If the null hypothesis is true, then, the probability of the obtained cell frequencies, within that set of events in which m units have been drawn from Population I and n units from Population II, is the product of two binomial probabilities, being

$\binom{m}{a} p^{a} (1-p)^{b} \binom{n}{c} p^{c} (1-p)^{d}$ or $\binom{m}{a} \binom{n}{c} p^{r} (1-p)^{s}$. The probability that of the N units in samples I and II combined, r will fall in category A and s in category B is $\binom{N}{r} p^r (1-p)^s$. Therefore the probability of the obtained cell frequencies within that set of N events in which m and n units are drawn from the respective populations, I and II, and r and s units fall into the respective column categories, A and B,

is $\dfrac{\binom{m}{a}\binom{n}{c} p^r (1-p)^s}{\binom{N}{r} p^r (1-p)^s}$. Since the unknown proportion, p, cancels

out, the probability of exactly the obtained cell frequencies with completely specified marginal frequencies m, n, r and s as shown is

$$\frac{m!\ n!\ r!\ s!}{N!\ a!\ b!\ c!\ d!}.$$

Since marginal frequencies are constants, this probability can be expressed in terms of a single cell frequency, becoming

$$\frac{m!\ n!\ r!\ s!}{N!\ a!\ (m-a)!\ (r-a)!\ (n-r+a)!}.$$ This is the probability for exactly the set of cell frequencies obtained, i.e., it is a point probability. The probability required, however, is the cumulative probability for those sets of cell frequencies which cause the greatest difference between the proportions a/m and c/n. Therefore the probability

$$\frac{m!\ n!\ r!\ s!}{N!\ a!\ (m-a)!\ (r-a)!\ (n-r+a)!}$$ must be cumulated over those values of a causing differences between the proportions a/m and c/n as great as or greater than that existing in the obtained table. If this cumulated probability is less than, α, the significance level chosen, the null hypothesis is rejected.

b. Null Hypothesis. The proportion of A's in Population I is the same as the proportion of A's in Population II.

c. Assumptions. (1) Sampling is random, (2) the N units are independent, i.e., to what categories a unit will belong is uninfluenced by the categories to which any other unit belongs (This assumption applies to the generation of the "table" and its marginal frequencies, and therefore is not in conflict with the fact that the table is completely specified by its marginal frequencies and a single cell frequency.), (3) the two row categories are mutually exclusive as are the two column categories, (4) the "A or B" dichotomy represents all possible "column" outcomes and the "I or II" dichotomy, all "row" outcomes (or, alternatively, sampling and statistical inference are restricted to that set of units capable of being dichotomized A or B in regard to one measured characteristic and I or II in regard to another). These assumptions are directly related to the assumptions of the binomials used in the derivation of the test. The assumption of independence is also occasioned by the fact that the probability for the obtained table was obtained by taking the product of the separate probabilities for the results in eachrow.

d. Efficiency and Power. In a sense the test is perfectly "efficient" since it is an exact method which uses all of the "information" in the sample; parametric tests for the same problem merely substitute the normal approximation for the true binomial distribution of frequencies within a cell and therefore use the same "information" but use it somewhat inaccurately. In the practical, computational sense, the test is inefficient for moderate and large samples if computation must be carried out without the aid of tables. Such tables do, however, exist for small and moderate size samples so the test may be regarded as practically inefficient only for application to large samples.

e. Application. To illustrate the application of this test, suppose that an experimenter has obtained the frequency data shown in the table below and wishes to test whether the true survival rate of persons afflicted by a rare disease is the same for men as for women.

	Survived	Died	
Men	4	10	14
Women	9	1	10
	13	11	24

The proportion of men surviving is 4/14 or .2857 while that for women is 9/10 or .90, and the difference between the two obtained proportions is .6143. Tables, with the same marginal totals, in which the difference between the proportions surviving is as great or greater than .6143 are shown below.

3	11
10	0

13	1
0	10

12	2
1	9

The values of a which cause the sex difference in the proportion surviving to be as great or greater than that in the obtained table are 3, 4, 12 and 13. Therefore the chance probability for results as extreme as those obtained, if there actually is no sex-fatality rate interaction is $\Sigma \frac{m!\ n!\ r!\ s!}{N!\ a!\ (m-a)!\ (r-a)!\ (n-r+a)!}$ with the summation being taken over the values a = 3, 4, 12 and 13 for a two tailed test. This probability is .00226. For the one-sided hypothesis that the survival rate for men is either greater than or equal to that for women, the summation is taken over a = 3 and a = 4 which gives a probability of .00208, i.e., which is "significant" at the .00208 level for a one-tailed test. For the opposite hypothesis that the survival rate for men is either less than or equal to that for women, the summation would be taken over the values a = 13, 12, 11, 10, etc., until the cumulative probability, on the next addition, would have exceeded the one-tailed significance level. Obviously for ordinary significance levels, this point would be reached before the probability for a = 4 was required in the summation, and since the critical region did not include the actually obtained value, a = 4, the hypothesis could not be rejected.

f. Discussion. The propriety of Fisher's Exact Method has been the subject of animated controversy among distinguished statisticians (2, 5, 14, 17, 29, 30, 38, 45). Some have objected that a test which necessarily takes marginal totals as fixed is therefore a "conditional" test and cannot properly be used as a basis for statistical inference to a larger, unrestricted population. The principle against which these objections were raised has subsequently become the basis of a number of distribution-free tests. It is that if two samples of sizes m and n have been drawn from identical populations, they may be regarded as a single random sample of size m+n from the common population. The two original samples may therefore be regarded as having been obtained by randomly assigning the label, "Sample I" to m of the m+n units in the "combined" sample, the n remaining units being labeled "Sample II". The degree to which Samples I and II differ, in any specified measure, not directly related to size, is therefore a matter of chance. The chance probability of the observed difference can therefore be obtained by forming all of the $\binom{m+n}{n}$ different possible "splits" of the common sample of m+n units into two samples of sizes m and n and by determining in what proportion of them the specified measure differs by an amount as great or greater than that actually obtained.

Applying this approach to the fourfold table, the marginal totals r and s may be regarded as the "parent" sample. There are $\binom{r+s}{m}$ or $\binom{N}{m}$ ways of splitting this sample into two samples of m and n units. The frequencies a and c can only be obtained from r and there are $\binom{r}{a}$ ways in which precisely these frequencies can be obtained for Samples I and II respectively. For each such way, there are $\binom{s}{b}$ ways of obtaining the frequencies b and d. Thus the point probability of the obtained table, given its marginal totals, is $\dfrac{\binom{r}{a}\binom{s}{b}}{\binom{N}{m}}$ or $\dfrac{m!\, n!\, r!\, s!}{N!\, a!\, b!\, c!\, d!}$, and the cumulative probability is obtained by summing the point probabilities for the appropriate values of a.

A number of different kinds of data can logically be cast into a fourfold table and a variety of hypotheses concerning the data are possible. Furthermore, the validity of a given hypothesis can be tested by a number of methods, although perhaps varying considerably in efficiency and logical appeal. These points have been made by critics of the method (2, 5, 19, 29). However, the Exact Method appears to be impeccable when used to test the null hypothesis that the unknown proportion of A's in two independent populations, capable of being dichotomized into mutually exclusive categories, A and B, is the same, and when the sample sizes m and n are determined in advance of sampling.

Unless samples are of equal size, the probability of a will not be the same as the probability of the "opposite deviation", m-a, in the upper left cell. Therefore, although when m=n two tailed probabilities can be obtained by doubling

$$\sum_{x=0}^{a} \frac{m!\; n!\; r!\; s!}{N!\; x!\; (m-x)!\; (r-x)!\; (n-r+x)!} \quad \text{or}$$

$$\sum_{x=a}^{m} \frac{m!\; n!\; r!\; s!}{N!\; x!\; (m-x)!\; (r-x)!\; (n-r+x)!} , \quad \text{whichever is smaller, when}$$

samples are of unequal size the summation must be taken over those extreme values of a which cause the absolute difference $|a/m - c/n|$ to be as great or greater than is the case in the actually obtained table. Furthermore, since a is an integer its probability is discretely distributed and there is not likely to be correspondence between integral values of a and the standard significance levels, .05, .01, and .001. If the experimenter has some compelling reason for wishing to use these standard significance levels he may employ Tocher's (38) modification: If the obtained cumulative probability is less than the standard significance level, α, results are considered significant at the level α. If the obtained cumulative probability exceeds α but would be less than α if cumulated for a value of a one unit more extreme, ("more extreme" values being understood to be those causing a larger absolute difference $|a/m - c/n|$), Tocher computes the ratio

$$\frac{\alpha - \text{Pr (more extreme a's)}}{\text{Pr (observed a or more extreme a's)}}.$$

He then enters a table of random numbers running from 0 to 1 and randomly selects a number. If the number selected is smaller than the above ratio, results are considered to have fallen within the α level of significance and the null hypothesis is rejected.

g. Tables. A number of tables (12, 13, 20, 21, 22, 23, 24, 42) have been prepared expressly for use with Fisher's Exact Method. Some have used Fisher's Exact Method to calculate probabilities when N is small, but have resorted to chi square with Yates' correction when N exceeds a certain value. In some of the tables it is suggested that two-tailed probabilities can be obtained by doubling the one-tailed probability listed in the table. This, of course, is strictly legitimate only if the distribution of the test statistic is symmetrical which, in fact, is the case only when the two samples are of equal size.

When N is small or when the significance level is extreme, probabilities may be obtained by a method described by Mosteller (II-34). The point probability of a is

$$\frac{[\binom{m}{a} p^a (1-p)^{m-a}][\binom{n}{c} p^c (1-p)^{n-c}]}{[\binom{N}{r} p^r (1-p)^{N-r}]}.$$

Each of the bracketed expressions is a binomial probability, and since the terms involving p cancel out, p may be arbitrarily assigned any constant value and the bracketed probabilities can then be obtained from tables of the point binomial. Thus the point probabilities of the most extreme values of a can be calculated and then cumulated.

h. Sources. 2, 5, 9, 12, 13, 14, 15, 17, 18, 20, 21, 22, 23, 24, 29, 38, 42, 44, 45, 46, 47. See also: 1, 3, 6, 16, 19, 28, 30, 31, 32, 33, 34, 36, 39, 48.

2. Westenberg's Median Test

a. *Rationale.* Let two samples of measurements be taken, one from Pupulation I, the other from Population II and let M be the median measurement of the pooled samples. If a and c are the respective numbers of measurements in Samples I and II which exceed M and if b and d are the corresponding numbers of measurements which are less than M, the data can be arranged in a fourfold table as follows and Fisher's Exact Method can be used to determine the probability that the proportion of measurements in Population I which exceed M is the same as the proportion of measurements greater than M in Population II.

	Above M	Below M	
Sample I	a	b	m
Sample II	c	d	n
	N/2	N/2	N

If the pooled sample median M be regarded as an *estimate* of the pooled population median, the test can be used to *test the hypothesis* that Populations I and II have identical medians. Otherwise it simply tests whether the value M splits Populations I and II into the same, but unknown, proportions.

b. *Null Hypothesis.* The proportion of measurements which lie above the median of Samples I and II combined is the same for Population I as for Population II.

A sufficient, but not a necessary, condition for the validity of the null hypothesis is that Populations I and II be identical. Therefore rejection of the null hypothesis is equivalent to rejection of the hypothesis of identical populations, but failure to reject the null hypothesis is *not* equivalent to failure to reject the hypothesis of identical populations.

c. *Assumptions.* As does Fisher's Exact Method, the test assumes *random sampling*, *dichotomized* and *mutually exclusive categories* for both rows and columns, and assumes that, in the process of sampling, each measurement value is *independent* of the value of every other measure (even though measurements are not independent in their a posteriori categorization). The median test assumes further that both populations are *continuously distributed* so that no measurements will be tied with M.

d. Treatment of Ties. Tied scores are a problem only when tied with M. If such ties constitute only a small proportion of N, half of the scores in each sample which are tied with M may be categorized as "above M", half as "below M". If, in a given sample there are an odd number of such ties, the odd tie may be discarded and the sample size reduced by one, or the odd tie may be categorized in whichever way will be least conducive to rejection of the null hypothesis. For a more conservative test, all scores tied with M may be categorized in the manner least conducive to rejection.

e. Efficiency. The asymptotic efficiency of the median test for location relative to Student's t-test, when both tests are applied to normal populations with equal variances, was found by Mood (V-37) to be $2/\pi$ or .637. Mood qualified his findings as resting upon certain unproved assumptions. Dixon (XI-13) found the power efficiency of the test, when sample sizes are very small, to be inferior to that of the Wilcoxon test and to that of the Maximum Absolute Deviation test when all three tests were applied to test the difference in means of two samples drawn from normal populations with equal variances. Lehmann (I-31) examined the relative power of six non-parametric tests when based on two small samples of equal size from two quite different continuous distributions. Ranked in order of decreasing power the tests were: Lehmann's "Most Powerful" test for the specific situation tested (one-tailed test), the Mann-Whitney test (one-tailed), Westenberg's Median test (one-tailed), the Mann-Whitney test (two-tailed), Westenberg's Median test (two-tailed), and finally the Wald-Wolfowitz Total Number of Runs test. Roughly, the median test was about 75% as powerful as the Mann-Whitney test. Apparently on the basis of these and his own results, Van der Waerden (I-52) concludes that the median test generally is less powerful than his X test.

f. Application. Fix sample size in advance and draw a sample from each of the two populations. Find the median, M, of the two samples when pooled, then determine the number of scores, a, in Sample I which are above, and the number, b, which are below M, counting half of the scores tied with M as "above", half as "below", and discarding any odd tie. Find the corresponding numbers, c, and, d, for Sample II, then construct the frequency table shown in "Rationale" with $m = a+b$, $n = c+d$ and $N = m+n$. Under this procedure, the frequency data entered in the fourfold table does not include the median score M, and the cell and marginal frequencies do not represent any discarded odd ties. From this point on, application is the same as for Fisher's Exact Method.

g. Discussion. The hypothesis actually tested is that equal proportions of Populations I and II lie above, and equal proportions lie below, the pooled sample median. If the pooled sample median, M, were the same value as the median of the pooled populations, the test would test whether or not Populations I and II had identical medians. However, this is almost certain not to be the case. When N is small the pooled sample median and the pooled population median may differ quite appreciably; for large values of N, however, the difference can be expected to be relatively small. Phrased differently, the median test tests whether or not the value, M, represents the same, but unknown, quantile in the two populations. If the null hypothesis is true and N is large this unknown quantile will be a proportion very close to .5 and the score M will be very nearly the common median of the two populations. The median test can, in this case, be regarded in an approximate sense as a test for identical population medians. However, when N is small, the validity of the null hypothesis does not insure that the unknown quantile represented by M will be in the neighborhood of .5, and the test can only be considered as testing whether the distributions of the two populations, when cumulated up to the point M, contain equal areas. If the two populations are identical this will be the case, so the small-sample median test can be used to test the hypothesis of identical population distributions. (See "Null Hypothesis").

If N is an odd number, the pooled sample median has the same value as one of the obtained scores. Since this score is neither above nor below M, it represents a third "binomial" outcome and violates one of the assumptions on which the test is based. (If N is large the consequence of this violation will be slight.) If N is even, this problem does not arise. However, in this case, M does not have a specific value, but rather can be defined only as lying somewhere between the $\frac{N}{2}$th and the $\frac{N}{2}+1^{th}$ ranked scores. Thus the null hypothesis, that equal proportions of the two populations lie above M, becomes equally vague. Summarizing, then, the median test is an approximate test for identical but unknown quantiles. As sample size increases, it becomes more nearly exact and the unknown quantile approaches .50 so that it tends to become a test for equal population medians.

h. Tables. All tables for Fisher's Exact Method are appropriate. (See 1. Fisher's Exact Method, g). Tables especially designed for median test have been published by Westenberg (40, 41, 43).

i. Sources. 26, 40, 41, 43.

3. The Median Test for Linear Trend

Cox and Stuart (11) have pointed out that if Sample I is taken to be the first half, and Sample II the second half, of a series of observations taken sequentially, the median test can be used to test for linear trend. If as time passes the population distribution, without changing in shape, simply "slides" upward or slides downward unidirectionally on the "x-axis", then the proportion of values above M in Population I will not be the same as the corresponding proportion in Population II. (Here Population I is the temporally changing population considered as existing from the beginning of sampling until half of the observations have been taken, Population II being similarly defined for the remaining interval.) And this statement will be equally valid whatever quantile M represents when the null hypothesis is true. Therefore, if it can be legitimately assumed that the sampled population may change in the location but not in the shape of its distribution, the test will be sensitive to "slippage" of any location parameter, and the question of how closely M represents the common population median will not be a problem.

Generally, however, a change in location is accompanied by a change in dispersion, and therefore by a change in the form of the population distribution. Therefore, in the generality of cases the additional assumption will not be legitimate. In such cases if the null hypothesis is false, the true, i.e., "alternative", hypothesis is that M is a different quantile in Population II than in Population I, i.e., the cumulative distributions of Populations I and II have different ordinates at the abscissa point M. If the additional assumption cannot be made, then, the test, in effect, tests for shift in an unknown quantile which may be near to or far from the population median.

The asymptotic relative efficiency of the median test for trend, relative to "the best (parametric) test against normal regression, based on the sample regression coefficient, b," is .78 (11, 35). This is the same as the A. R. E. of Cox and Stuart's S_2 sign test for trend.

4. Westenberg's Test for Interquartile Range

Westenberg (43) has proposed a modification of his own median test in which, instead of dividing each sample into observations above and observations below the median of the pooled sample, the samples are divided into observations within and observations outside of the interquartile range, Q_1 to Q_3, of the pooled sample.

	Within $Q_1 - Q_3$	Outside $Q_1 - Q_3$	
Sample I	a	b	m
Sample II	c	d	n
Total	N/2	N/2	N

Since the expected proportion of observations above a median is the same as the expected proportion of observations within an interquartile range, the two tests have identical mathematical bases. The performance of the interquartile range test is therefore analogous to that of the median test. The null hypothesis is that identical proportions of Populations I and II lie within the interquartile range of the pooled samples. The test therefore does not test whether the two populations have equal interquartile ranges; it tests whether they have equal areas included between the values Q_1 and Q_3 which were obtained from the *samples.* (See "Discussion" of the median test.) The efficiency of the test apparently is unknown. Treatment of ties is analogous to that of the median test; all ties may be categorized conservatively; or in each sample, half of the observations tied with either Q_1 or Q_3 be counted as "within", half as "outside" and any odd tied observation discarded.

5. A "Median" Test for Correlation

a. *Rationale.* Consider a sample of units or individuals upon each of which an x measurement and a y measurement have been made. Let its scattergram be divided into four quadrants by a horizontal line through the sample's y median and a vertical line through its x median. Then if the x and y attributes are uncorrelated, one would expect each of the four quadrants to contain about the same number of units; while, if a correlation exists, a preponderance of units should be located in one of the two pairs of diagonal quadrants.

If the x and y attributes are uncorrelated, dividing the original sample into two equal sized samples on the basis of some characteristic of x will divide the y's into two "y-samples" which differ on the basis of chance alone. They are therefore two samples from the same population of y's, and in each sample the proportion of y's having any specified y characteristic should also differ on the basis of chance alone. On the other hand if x and y are correlated in respect to the criteria used to subdivide the sample, the two y-samples will, in a sense, be from different populations which contain different proportions of y's with the relevant , specified characteristic.

This treatment of correlation reduces therefore to Fisher's Exact Method with categories as shown below:

	A:	B:	
	Above	Below	
	Sample	Sample	
	y median	y median	Total
Sample I: y's whose paired x is above sample x median	a	b	m
Sample II: y's whose paired x is below sample x median	c	d	n
Total	r	s	N

The categorizations and designation of table frequencies can be simplified to the following, "units" being the item tabled, a unit's x measure being referred to in the rows, its y measure in the columns.

	Above	Below	
	y median	y median	Total
Above x median	a	$\frac{N}{2} - a$	N/2
Below x median	$\frac{N}{2} - a$	a	N/2
Total	N/2	N/2	N

The point probability for the tabled frequencies, if the null hypothesis of no correlation is true, is therefore

$$\frac{[(N/2)!]^4}{N!\ (a!)^2\ [(\frac{N}{2} - a)!]^2} .$$

b. Null Hypothesis. In the parent population, those units whose x value exceeds the sample x-median have the same proportion of y's above the sample y-median as have those units whose x value is less than the sample x-median. A sufficient condition for its validity is that the x and y attributes are uncorrelated.

c. Assumptions. Same as for Westenberg's median test; see 2.

d. Treatment of Ties. Tied scores are a problem only when tied with one or both of the sample medians. For a conservative test all such ties may be categorized in the manner least conducive to rejection of the null hypothesis. Alternatively, to minimize tie error, half of the scattergram units lying on the line separating two quadrants may be counted as belonging to each quadrant. If there are an odd number of such units, the odd unit should be held for discarding. Units lying on the intersection of the two median lines should be discarded. Before discarding, a certain number of "units" may be salvaged. For example, if one unit has its x value tied with the x median and another unit has its y value tied with the y median, two new "units" may be formed from the old ones, one of which has nontied x and y values, the other having both x and y values tied with their medians. Only the latter new unit need be discarded, the former being "returned" to the sample. The value N should refer to the number of units remaining in the sample after all discarding has been completed. When ties are treated in this manner, marginal frequencies need not all equal N/2 so the formula

$\frac{m!\ n!\ r!\ s!}{N!\ a!\ b!\ c!\ d!}$ from Fisher's Exact Method should be used to calculate probabilities.

e. Efficiency. Applied to populations known to have normally distributed x's and normally distributed y's, the test has an asymptotic local efficiency of $(2/\pi)^2$ or .41 relative to the correlation coefficient ρ. Under the same circumstances its asymptotic efficiency relative to Kendall's rank order correlation coefficient, τ, is 4/9. (4)

f. *Application.* Find the sample x and y medians, construct the fourfold table shown in "Rationale" treating ties as outlined under (d), and apply Fisher's Exact Method.

g. *Discussion.* If no correlation exists, then, on the average, half of the sample units should fall in the "North-West" and "South-East" quadrants and half in the opposite diagonal pair. It might be supposed, therefore, that the number of units, r, in one of the pairs of diagonal quadrants would be binomially distributed with $p = .50$ when the null hypothesis is true, so that $\binom{N}{r} (.50)^N$ would be the point probability of the obtained results. Such a supposition would be in error. The binomial test would require that the categorization of each unit to one of the diagonal pairs of quadrants be *independent* of the categorization of every other unit. However, it is in the nature of the construction of the table that equal numbers of units must "fall" in diagonally opposite quadrants. Thus, for each unit falling in a given quadrant, another unit *must* fall in the diagonally opposite quadrant and therefore *must* receive the *same* binomial categorization given the first unit. For example, if N = 4, there are three possible tables: $\begin{array}{c|c} 0 & 2 \\ \hline 2 & 0 \end{array}$, $\begin{array}{c|c} 1 & 1 \\ \hline 1 & 1 \end{array}$, and $\begin{array}{c|c} 2 & 0 \\ \hline 0 & 2 \end{array}$. There are 6 permutations of the N units which will give the first table, 24 which will yield the second, and 6 which result in the third. Thus the respective probabilities of the three tables are 6/36, 24/36 and 6/36 or 1/6, 4/6 and 1/6. These are also the probabilities obtained by using the formula $\dfrac{[(N/2)!]^4}{N!\,(a!)^2\,[(\frac{N}{2} - a)!]^2}$.

If the binomial test is applied to the three tables, the respective probabilities are calculated to be 1/16, 6/16, and 1/16. Not only are these "probabilities" different, but their sum is 1/2 rather than 1, clearly indicating that the test is fundamentally in error. The sum of the "probabilities" is 1/2 rather than 1 because the number of units in a pair of diagonally opposite quadrants can only be an even number, while a truly binomial variate can assume any integral value between zero and N. Nor would it be correct to confine the binomial test to, say, the upper two quadrants, calculating the probability that, of the N/2 units in the upper two quadrants a of them would fall in the left quadrant. In the upper half of the tables just discussed, the number of permutations of the N/2 units which will give the three results shown are 1, 2, and 1. The probabilities for the upper halves of the three tables, considered separately and as if independent of the lower halves, are therefore 1/4, 2/4, and 1/4, which are also those obtained by using the binomial formula. Thus the table, taken as a whole, has a different probability than its upper half alone. Clearly, then the dependence between units in diagonally opposite quadrants is a partial dependence which can neither be ignored, by applying a binomial test to the number of units in a diagonal pair of quadrants, nor be treated as a complete dependence by confining the binomial test to the upper half of the table. The error shown to exist in the binomial approach is not confined to very small sample sizes. For example, the table $\begin{array}{c|c} 0 & 8 \\ \hline 8 & 0 \end{array}$ has the "probabilities", 1/12,870, 1/65,536, and 1/256 respectively when tested by Fisher's Exact Method, by the binomial test applied to the entire table, and by the binomial test applied only to the upper half of the table.

If the sample is divided into quadrants by its x and y means, rather than medians, the "binomial" approach is still unconscionable. If median and mean are identical all of the objections discussed above apply. If they differ, the premise that half of the sample units would be expected to lie in a pair of diagonally opposite quadrants is false, and the binomial parameter, p, does not have the value, .50, substituted in the formula used to calculate probabilities.

h. Tables. Tables for Fisher's Exact Method are appropriate. See 1.

i. Sources. 4, 8, 10.

6. Test for a Difference between Correlated Proportions

a. Rationale. If each of N units or individuals have been categorized as belonging to one or the other of two mutually exclusive categories I and II, and the same N units have been categorized according to another mutually exclusive dichotomy A and B, the experimenter may wish to know whether or not in the parent population the proportion of I's differs from the Proportion of A's. Let the frequency data be represented by the accompanying table.

	A	B	
I	a	b	m
II	c	d	n
	r	s	N

Letting primes indicate population values corresponding to sample frequencies, the proportion of I's is m'/N' or $\frac{a' + b'}{N'}$ and the proportion of A's is r'/N' or $\frac{a' + c'}{N'}$. These two proportions are equal only if $b' = c'$. Therefore, the hypothesis of equal proportions can be tested by examining the probability of obtaining the sample b and c by random sampling of b+c units from an infinite population consisting of equal numbers of b'' s and c'' s. Thus the point probability for the obtained b and c is given by the binomial $\binom{b+c}{b} (.5)^{b+c}$.

b. Null Hypothesis. In an infinite population of units each of which is classed as either I or II and as either A or B, the proportion of units categorized as I's has the same value as the proportion of units categorized as A's. If this hypothesis is true, it follows inevitably that there are exactly as many II A units as I B units in the population of I A's, II A's, I B's and II B's, and this is the hypothesis actually tested.

c. Assumptions. Since the test is a binomial one, it depends upon the usual binomial assumptions: (1) sampling is random, (2) categorization of one unit does not influence the categorization of any other unit, i.e., units are independent and are drawn from an infinite population of potential units, (3) the population selected for test, i.e., the units categorized II A or IB, constitutes a dichotomy, (4) the dichotomized categories II A and I B are mutually exclusive. In addition, the unique construction of the test necessitates the following assumptions: (5) the I and II categories are mutually exclusive as are the A and B categories, thereby making the four categories I A, II A, I B, and II B mutually exclusive (the latter is required in order that the "no trial" categories, I A and II B, will contain none of the II A or I B attributes, those actually tested, thus making the exclusion of I A and II B data legitimate.) (6) every unit categorized either I or II is also categorized either A or B and vice versa, i.e., the "I II" and "A B" categorizations are applied to the same data; unless this is the case, the data cannot legitimately be cast into a fourfold table, but specifically the proportions of I's and A's cannot be represented as $\frac{a' + b'}{N'}$ and $\frac{a' + c'}{N'}$ respectively, and a difference between b' and c' is not sufficient to demonstrate a difference between the two proportions.

d. Efficiency. No information seems to be available; however, it would appear logical that the test efficiency would be high since the test appears to make efficient use of all the "information" available.

e. Application. Draw a sample of N units from the population in question, and let the table shown in "Rationale" represent the frequency data categorized according to each of the dichotomies I or II and A or B. Let α represent the level of significance chosen, and let r represent the smaller of the two frequencies, b and c.

For a two-tailed test of the null hypothesis that in the parent population the unknown proportion of I's is the same as the correlated, unknown proportion of A's, reject if $2 \sum_{i=0}^{r} \binom{b+c}{i} (.5)^{b+c} \leq \alpha$. For a one-tailed test, reject the hypothesis that the proportion of I's is either the same or smaller than the proportion of A's if $\sum_{i=0}^{c} \binom{b+c}{i} (.5)^{b+c} \leq \alpha$. Or, for the opposite one-tailed test, reject the hypothesis that the proportion of I's is either the same or greater than the proportion of A's if $\sum_{i=0}^{b} \binom{b+c}{i} (.5)^{b+c} \leq \alpha$.

f. Discussion. McNemar (25), who originated the test, used the chi square approximation, rather than the binomial, with

$$\chi^2 = \frac{\left(b - \frac{b+c}{2}\right)^2}{\frac{b+c}{2}} + \frac{\left(c - \frac{b+c}{2}\right)^2}{\frac{b+c}{2}} \text{ which reduces to } \chi^2 = \frac{(b-c)^2}{b+c}$$

with one degree of freedom. The binomial, however, is the exact test and should be used unless b+c is very large, in which case either test may be used.

Although this test bears a superficial similarity to the binomial test for correlation criticised in the "Discussion" section of (5. A "Median" Test for Correlation), the objections voiced there do not apply here. In the present test, categories are completely specified in advance of sampling, the categorization of one unit does not influence the categorization of any other unit, and the "population" from which the sample is considered to have been obtained is the parent population from which the b+c units were drawn. In the binomial test for correlation, on the other hand, categories were established after sampling and were a function of the sample results, and the categorizations of units were not independent. The proper analysis of such data requires that the test be a "conditional" test in which the obtained *table* is regarded as a sample from a population of tables. Each table in a population of tables with fixed marginal frequencies is a different permutation of the units constituting the cell frequencies. Therefore, in calculating probabilities for such conditional tests all permutations and therefore all cells must be considered. Since McNemar's test is not a conditional test, no restrictions having been placed on marginal frequencies, units may distribute themselves in the "b" and "c" cells strictly according to the binomial law.

g. *Tables.* Use tables of the cumulative binomial probability with $p = .5$, or tables for the Sign Test. (See Chapter II). Tables especially designed for the application of this test employing the chi square approximation to the binomial have been published by Swineford (37).

h. *Sources.* 25, 37.

BIBLIOGRAPHY

* 1. Barnard, G. A., A new test for 2 X 2 tables. Nature, London, 1945, 156, 177.

2. BARNARD, G. A., Significance tests for 2 X 2 tables. Biometrika, 1947, 34, 123-138.

3. Barnard, G. A., 2 X 2 tables. A note on E. S. Pearson's paper. Biometrika, 1947, 34, 168-169.

* 4. Blomqvist, N., On a measure of dependence between two random variables. Annals of Mathematical Statistics, 1950, 21, 593-600.

5. Bonnier, G., The four-fold table and the heterogeneity test. Science, 1942, 96, 13-14.

6. Bross, I., Misclassification in 2 X 2 tables. Biometrics 1954, 10, 478-486.

7. Brown, G. W. and Mood, A. M., On median tests for linear hypotheses., Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. (Ed. by Jerzy Neyman), University of California Press, 1951, 159-166.

8. Chown, L. N. and Moran, P. A. P., Rapid methods for estimating correlation coefficients. Biometrika, 1951, 38, 464-467.

9. Cochran, W. G., The X^2 test of goodness of fit. Annals of Mathematical Statistics, 1952, 23, 315-345.

10. Cochran, W. G., The efficiencies of the binomial series test of significance of a mean and of a correlation coefficient. Journal of the Royal Statistical Society, 1937, 100, 69-73.

* 11. COX, D. R. and STUART, A., Some quick sign tests for trend in location and dispersion. Biometrika, 1955, 42, 80-95.

T 12. Federighi, E., The use of chi-square in small samples. American Sociological Review, 1950, 15, 777-779.

T 13. Finney, D. J., The Fisher-Yates test of significance in 2 X 2 contingency tables. _Biometrika,_ 1948, 35, 145-156.

14. Fisher, R. A., A new test for 2 X 2 tables. _Nature,_ London, 1945, 156, 388.

* 15. Fisher, R. A., _Statistical methods for research workers,_ London: Oliver and Boyd, 1946, 96-97.

16. Fisher, R. A., Statistical tests of agreement between observation and hypothesis. _Economica,_ 1923, 8, 1-9.

17. Fisher, R. A., The interpretation of experimental four-fold tables. _Science,_ 1941, 94, 210-211.

18. Fisher, R. A., The logic of inductive inference. _Journal of the Royal Statistical Society,_ 1935, 98, 39-82.

19. Irwin, J. O., Tests of significance for differences between percentages based on small numbers. _Metron,_ 1935, 12, 83-94.

T 20. Latscha, R., Tests of significance in a 2 X 2 contingency table: Extension of Finney's table. _Biometrika,_ 1953, 40, 74-86.

TT 21. Mainland, D., Statistical methods in medical research I. Qualitative statistics (enumeration data). _Canadian Journal of Research,_ Sec. E, 1948, 26, 1-166.

TT 22. MAINLAND, D., HERRERA, L. and SUTCLIFFE, MARION, _Tables for use with binomial samples,_ New York: Dept. Medical Statistics, New York University College of Medicine, 1956.

T 23. Mainland, D. and Murray, I. M., Tables for use in fourfold contingency tests. _Science,_ 1952, 116, 591-594.

T 24. Mainland, D. and Sutcliffe, Marion, Statistical methods in medical research II Sample sizes required in experiments involving all-or-none responses. _Canadian Journal of Medical Science,_ 1953, 31, 406-416.

* 25. McNemar, Q., Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 1947, 12, 153-157.

26. Mood. A. M., Introduction to the theory of statistics, New York: McGraw-Hill, 1950, 394-395.

27. Moore, P. G., A test for randomness in a sequence of two alternatives involving a 2 X 2 table. Biometrika, 1949, 36, 305-316.

28. Patnaik, P. B., The power function of the test for the difference between two proportions in a 2 X 2 table. Biometrika, 1948, 35, 157-175.

29. Pearson, E. S., The choice of statistical tests illustrated on the interpretation of data classed in a 2 X 2 table. Biometrika, 1947, 34, 139-167.

30. Pearson, E. S. and Merrington, Maxine, 2 X 2 tables; the power function of the test on a randomized experiment. Biometrika, 1948, 35, 331-345.

31. Pearson, K., On the difference and double tests for ascertaining whether two samples have been drawn from the same population. Biometrika, 1924, 16, 249-252.

32. Rhodes, E. C., On the problem whether two given samples can be supposed to have been drawn from the same population. Biometrika, 1924, 16, 239-248.

33. Sillitto, G. P., Note on approximations to the power function of the '2 X 2 comparative trial'. Biometrika, 1949, 36, 347-352.

34. Stevens, W. L., Mean and variance of an entry in a contingency table. Biometrika, 1951, 38, 468-470.

35. Stuart, A., The efficiencies of tests of randomness against normal regression. Journal of the American Statistical Association, 1956, 51, 285-287.

T 36. Swaroop, S., Tables of the exact values of probabilities for testing the significance of differences between proportions based on pairs of small samples. Sankhyā, 1938, 4, 73-84.

T 37. Swineford, Frances, A table for estimating the significance of the difference between correlated percentages. Psychometrika, 1948, 13, 23-25.

38. Tocher, K. D., Extension of the Neyman-Pearson Theory of tests to discontinuous variates. Biometrika, 1950, 37, 130-144.

T 39. Trites, D. K., Graphic determination of significance of 2 X 2 contingency tables. Psychological Bulletin, 1957, 54, 140-144.

T 40. Westenberg, J., A tabulation of the median test for unequal samples. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1950, 53, 77-82.

41. Westenberg, J., A tabulation of the median test with comments and corrections to previous papers. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1952, 14, 10-15.

T 42. Westenberg, J., Mathematics of pollen diagrams I and II. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1947, 50, 509-520 and 640-648.

** 43. WESTENBERG, J., Significance test for median and interquartile range in samples from continuous populations of any form. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 1948, 51, 252-261.

44. Wilson, E. B., On contingency tables. Proceedings of the National Academy of Science (USA), 1942, 28, 94-100.

45. Wilson, E. B., The controlled experiment and the four-fold table. Science, 1941, 93, 557-560.

46. Wilson, E. B. and Worcester, Jane, Contingency tables. Proceedings of the National Academy of Science, (USA), 1942, 28, 378-384.

47. Yates, F., Contingency tables involving small numbers and the X^2 test. Journal of the Royal Statistical Society, (Series B), 1934, 1, 217-235.

48. Yule, G. U., On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 1912, 75, 579-652.

CHAPTER V

TESTS BASED ON FISHER'S METHOD OF RANDOMIZATION I

The logical basis for most distribution-free tests is rooted in a method originated by R. A. Fisher and known as the Method of Randomization. The basis of statistical inference is simply this. If several samples have been drawn from a common population, they may be regarded as one large sample whose observations have been randomly assigned to subsamples or component samples of the sizes actually drawn. Each of the different possible random assignments was, prior to sampling, equally likely to be the actually obtained sample, if the null hypothesis of identical populations is true, but unequally likely to be if the null hypothesis is false. By choosing a test statistic which is sensitive to the alternative hypothesis and calculating its value for each of the n different possible random assignments, one obtains a set of n equally weighted values of the test statistic (some of which are the same) which form the distribution of the test statistic under the null hypothesis. Its rejection region is simply the N most extreme of these values each of which is exactly as likely as any other value when the null hypothesis is true, but which become especially probable when the alternative hypothesis is true. If the test statistic for the actually obtained sample falls within the rejection region, the null hypothesis can be rejected at the N/n level of significance.

The method, as developed by Fisher, has been improved by Wilcoxon who, by replacing original observation magnitudes by their ranks, "standardized" the rejection region and permitted tabling of probabilities. Wilcoxon's tests are among the most efficient and most important distribution-free tests. The sample information used by Fisher was the sample mean or mean difference; Wilcoxon used rank sums or sums of algebraically signed ranks. Both constructed tests sensitive to location.

1. Fisher's Method of Randomization: Matched Pairs

a. Rationale. Let n matched pairs of observations be taken, one member of each pair having been taken under treatment A, the other under treatment B. If each B observation is subtracted from its paired A observation, there will be n difference scores henceforth referred to as the obtained sample. If the A and B treatments have equal effects, in all respects to which the measurements are sensitive, then the members of any given matched pair of observations may be regarded as having been drawn from the same population. In this case "treatment A" and "treatment B" are merely arbitrary labels which are applied to two random observations from the same population, and a specified one of the two observations is as likely to acquire the label "A" as to be labeled "B". The difference score for any given pair of observations is therefore as likely to be plus as to be minus. If the A and B treatments produce effects whose distributions are not identical but which are symmetrical about the same point, a given difference score is also as likely to be plus as to be minus because for each $A_i - B_i$ difference score in the population there is an equally likely "mirror-image" difference score of equal magnitude but opposite sign. Therefore if either (a) the A population and the B population are identical, or (b) if the A and B populations are symmetrical about the same point, each difference-score, whatever its magnitude, will be as likely to be plus as to be minus. Since plus and minus are equally likely algebraic signs for each of the n difference magnitudes, each of the 2^n different possible arbitrary assignments of algebraic signs to the obtained difference magnitudes is equally likely for a sample containing these difference magnitudes (provided no difference magnitudes are zero for which an algebraic sign is meaningless). That is to say, there are two ways of assigning algebraic sign to the first difference magnitude; for each of these ways there are two ways of assigning sign to the second magnitude, making four distinguishable combinations; for each of these four combinations the third magnitude can be treated in two ways, making eight combinations, etc., so that for n difference scores there are 2^n distinguishable patterns of algebraic sign which can be "superimposed" upon the obtained set of difference magnitudes; and if the sampled populations are either identical or symmetrical about the same point each of these 2^n sets of difference scores were exactly as likely to have been drawn as a sample as was the set constituting the obtained sample.

Imagine now that for each of the 2^n sets of difference scores a mean difference has been calculated by summing the n difference scores and dividing by n. If the A and B populations are identical or are symmetrical about the same point, each of these 2^n mean differences will be equally probable. The N largest of these 2^n mean differences should therefore contain the mean difference for the obtained sample in exactly a proportion $\frac{N}{2^n}$ of such experiments. On the other hand, if the A and B populations are identical in form but differ in location, or if they are both symmetrical but not symmetrical about the same point, the mean difference for the obtained sample is more likely to lie among the extreme N of the 2^n mean differences than the proportion $\frac{N}{2^n}$ would imply. And even if the two populations have nonidentical, asymmetrical forms, one would generally expect large mean differences to be more likely than small ones if the populations have different means.

b. Null Hypothesis. Each of the 2^n unique sets of difference scores obtainable by arbitrarily assigning algebraic signs to the obtained difference-score magnitudes is equally likely to have been drawn as a sample. Either of two conditions is sufficient to insure the validity of the null hypothesis: (a) the sampled populations are identical, (b) the sampled populations are both symmetrical and are symmetrical about a common point. By taking as the rejection region the N sets with the N greatest mean differences, the method of randomization tests the null hypothesis that populations are identical or symmetrical about a common point against the alternative that the populations have different means. It is merely "most sensitive" against this alternative, however, since nonidentity of populations with equal means can also cause rejection. Certain assumptions, therefore, are necessary to eliminate such alternatives when they are not desired.

c. Assumptions. By taking as the probability fraction the ratio of the number of ways certain events can occur, it is implied that each way is equally probable when the null hypothesis is true and unequally probable when it is false. However, they can be unequally probable, not because populations violate the null hypothesis, but rather because of bias in the selection of samples or because of the influence of one sample unit upon another. Therefore, in order to eliminate such contingencies, it is assumed that sampling is random and observations are independent.

By using 2^n as the denominator of the probability fraction, it is implied that each difference-score has two possible values, one plus and the other minus. This means that there must be no zero differences, or the equivalent, but more general, assumption of continuously distributed populations may be made.

If the populations do not have the same form or if they are not symmetrical, then the obtained difference scores are not necessarily as likely, a priori, to be minus as to be plus even though the sampled populations have equal means. In order therefore to "eliminate" such causes of unequally likely signs and confine the cause to unequal population means, it is necessary to introduce the assumption that either (a) the two sampled populations have identical forms, differing, if at all, only in location, or (b) each sampled population has a symmetrical distribution, the two distribution forms not necessarily being the same.

d. Treatment of Ties. If the number of zero differences, t, is small relative to the total number of difference scores, discard them and reduce n by t in all subsequent calculations, so that the denominator of the probability fraction is 2^{n-t}. It should be borne in mind that discarding the zero differences artificially increases the power of the test.

e. Efficiency. No figures appear to be available; however, there is reason to believe efficiency should be high. See Wilcoxon test.

f. Application. As an example, suppose that each of seven individuals have been subjected to each of two treatments, A and B, and that there are no sequential or interaction effects between treatments. The data are presented in the following table.

	SCORES							Σ	MEAN
Treatment A	23	16	11	12	9	5	1	77	11
Treatment B	8	5	2	7	6	4	3	35	5
Difference: A-B	15	11	9	5	3	1	-2	42	6

There are 2^7 or 128 different ways of distributing plus and minus signs among the seven difference scores. Three of these ways result in a positive mean difference, and six result in an absolute mean difference, as great or greater than that actually obtained. They are as follows:

Difference Scores							Σ	Mean
15	11	9	5	3	1	2	46	6.57
15	11	9	5	3	-1	2	44	6.29
15	11	9	5	3	1	-2	42	6.00
-15	-11	-9	-5	-3	-1	-2	-46	-6.57
-15	-11	-9	-5	-3	+1	-2	-44	-6.29
-15	-11	-9	-5	-3	-1	+2	-42	-6.00

As indicated, only a small number of the 2^n mean differences need actually be calculated, specifically those equal to or more extreme than that actually obtained or those constituting the rejection region, whichever is less. Therefore, assuming populations identical in form, the hypothesis that treatments have equal effects can be rejected at the 6/128 or .047 level of significance in favor of the alternative hypothesis that the mean effects of the two treatments differ. Or, under a one-tailed test the hypothesis that treatment A has the same effect or less mean effect than treatment B can be rejected at the 3/128 or .023 level of significance in favor of the hypothesis that treatment A has more mean effect than treatment B. If it can be assumed that populations are either identical in form or symmetrical , the term "effect" must be replaced by "mean effect" in the expression of the null hypothesis.

g. Discussion. The magnitudes of the n difference scores are, with rare exceptions, unequally likely. However, if the sampled populations are identical or symmetrical about a common point, each of the 2^n differently "signed" sets of difference scores is equally likely because each set contains the same magnitudes and each magnitude is as likely to be positive as to be negative. If the null hypothesis is false, one of the two algebraic signs will be more probable than the other. The more probable sign would be expected either to occur more frequently than its opposite, or to be associated more frequently with the larger than with the smaller magnitudes, or both. The likelihood that a difference score had the more probable algebraic sign would be expected to increase with the absolute magnitude of the difference score. By taking as the rejection region those sets of difference-scores (equally probable when the null hypothesis is true) which yield the most extreme mean differences, one is quite properly permitting the larger magnitudes to influence rejection more than the smaller ones. Thus each algebraic sign may be considered to be "weighted" by the difference-score magnitude to which it is attached. This weighting is arbitrary, i.e., randomly determined, when the null hypothesis is true and the distribution of the test statistic is such that each weight is applied as frequently to positive as.to negative signs. It is only when the null hypothesis is false that the weighting takes on a discriminating function, making the test especially sensitive to differences in location.

The sample space for the test statistic consists of the 2^n sets of difference scores obtainable by varying the signs attached to the same set of n difference- score magnitudes. The test is therefore a conditional test in the sense that the probability fraction $\frac{N}{2^n}$ gives the chance probability of drawing the obtained sample, or a more extreme one, from that artificially limited sample space rather than from the larger parent population of difference scores from which it was actually drawn. The importance of this fact has been frequently overemphasized. When the null hypothesis is true every difference score in the sampled population is as likely to be plus as to be minus, not just those in the restricted sample space. Therefore the probability of commiting a Type I error is unaffected by restricting the sample space, being exactly $\frac{N}{2^n}$ whatever the particular set of difference scores sampled. When the null hypothesis is false the relative probability of possession of the two algebraic signs may differ greatly from one population difference-score magnitude to another and not necessarily in any direct relationship to the absolute size of the magnitude. Since chance determines which of these population difference-scores will be drawn for the sample, chance plays a large role in determining whether or not a false hypothesis will be rejected. However, this is equally true of nonconditional tests. It is more or less assumed, for both conditional and nonconditional tests, that the sample is fairly representative of the population. To the extent that this is untrue both types of test are likely to err; to the extent that it is true the restriction of the sample space of Fisher's conditional test statistic is not a serious shortcoming of the test.

In connection with criticisms of the conditional nature of Fisher's test it has sometimes been fallaciously implied that the test statistic has the same distribution under an alternative hypothesis as it has under the null hypothesis. When the null hypothesis is false, just as <u>many</u> of the 2^n possible values of the test statistic lie in the rejection region as when the null hypothesis is true. However, when the null hypothesis is true each of these 2^n values is equally probable, whereas when it is false, those values occupying the rejection region are more probable than the ones occupying the acceptance region, thus biassing the test (properly so) in favor of rejection. Student's t test operates in much the same way. The set of possible values of t is the same whether the null hypothesis is true or false; it is only their probabilities which differ. When the null hypothesis is true the possible values of t constituting the rejection region have a cumulative probability of α, whereas when it is false they have a cumulative probability greater than α. It is incorrect, therefore, to imply, as has been done, that under the method of randomization the test statistic has the same distribution under alternative hypotheses as under the null hypothesis. This is no more true of the method of randomization than of Student's t. Although Fisher's and Student's tests operate in somewhat similar ways, however, Fisher's test cannot be regarded as giving the "true" probability which Student's test "approximates". This has sometimes been implied, the difference in the two probabilities being attributed to violations of the assumptions of Student's test or to other artifacts. The argument, however, is fallacious. The two tests cannot be expected to yield equal probabilities when applied to the same sample because (a) the test statistics do not have the same distribution, (b) the tests do not use the same rejection region.

Although many of the criticisms of the method of randomization have been overstated, it does have a number of shortcomings which will be outlined in the following paragraphs.

Two types of information are used in the test: algebraic sign and magnitude. When the null hypothesis is true magnitudes are randomly associated with equally likely algebraic signs. When it is false magnitudes become nonrandomly associated with unequally probable algebraic signs in a complex way: for some magnitudes one algebraic sign becomes more probable than the other, and for other magnitudes the reverse is probably the case. Presumably the larger the magnitude the more likely it usually is to have the algebraic sign indicating the true direction of difference; however, there is no justification for assuming that this relationship is linear or even monotonic. Since each sample consists of a different set of magnitudes and since the magnitudes are, in effect, weights, each sample from the same population is subjected to a different weight function. Since the weight function varies from sample to sample and since the relationship of weight to the probability of a given algebraic sign is unknown, probability levels for samples from the same population are not strictly comparable. Another way of stating this is that probability levels are not strictly comparable because no two samples use the same rejection region.

Another, related, disadvantage of Fisher's method is that the test is quite sensitive to isolated extreme difference scores. Suppose, for example, that the obtained set of difference-scores were +1, +2, +3, +4, +5, +6, +7, +8, +9, +50. There are $2^{10} = 1024$ possible ways of assigning signs to these magnitudes and the mean difference for the obtained sample can be equaled or exceeded in only one of them, so the obtained sample has a one-tailed probability of 1/1024 or less than .001. However if the algebraic sign of the 50 is changed to minus, the obtained mean difference becomes -.5 which can be exceeded by any of the 512 assignments in which the 50 is plus. The one-tailed probability therefore drops from less than .001 to slightly more than .50 simply by changing the sign of one of ten difference scores. This is in no way improper since, if the null hypothesis is false, one would expect the difference in probability between a +50 and a -50 to be much greater than the difference in probability between a+1 and a-1. However it shows that the test gives great weight to isolated extreme differences which frequently one wishes to deemphasize because of the likelihood that they are spurious or represent atypical performance (or response).

A final disadvantage is that Fisher's method of randomization requires that of the 2^n possible "ways" of calculating a mean difference (using the same set of n difference magnitudes but varying their algebraic signs) the experimenter must actually enumerate either the number of ways constituting the rejection region or the number of ways which result in a mean difference equaling or exceeding the one obtained, whichever is less. If n is large, or if n is of moderate size and α is large, the computations are likely to be so lengthy as to make the test impractical. Since the exact forms of the sampled populations are unknown the sample difference scores are of unpredictable magnitude and it is impossible to construct probability tables in advance of sampling.

h. Tables. None. Probabilities must be calculated for each specific case.

i. Sources. 4, 7, 17, 26, 27, 34, 38, 39, 40, 48, 75. See also 16, 28, 41, 42, 43, 67, 68 under 4, Fisher's Method of Randomization: Unmatched Data.

2. The Wilcoxon Test: Matched Pairs

a. Rationale. Wilcoxon has modified Fisher's method by replacing the obtained difference-scores with the ranks of their absolute magnitudes, each rank being given the algebraic sign of the difference-score which it replaces. The test statistic is the algebraic sum of the signed ranks rather than the average signed rank; since the former is always n times the latter, the two have equivalent disbributions. Wilcoxon's modification has several advantages over the original test. First, the test is not a conditional one since the sample space for the test statistic is the same for every sample. Thus every sample is made comparable with every other sample of the same size in the sense that the set of numbers by which the signs of the differences are weighted is always the same: the sign of the largest difference magnitude always being given a weight of n, the next largest, n-1, etc. Second, the test is less sensitive to extreme difference-score magnitudes since the most extreme magnitude will receive a rank only one greater than the next-to-extreme magnitude, etc. Finally, by using ranks, the probabilities can be tabled, since for any given n, instead of n random and unpredictable magnitudes, the magnitudes consist always of the integers 1 to n.

If each obtained difference-score magnitude is as likely to be plus as to be minus, then so is its rank. The rationale for the Wilcoxon test therefore parallels that for Fisher's method of randomization. See 1, Fisher's Method of Randomization: Matched Pairs.

b. Null Hypothesis. Each of the 2^n unique sets of signed ranks, obtainable by arbitrarily assigning algebraic signs to the ranks of the difference-score magnitudes from the obtained sample, is equally likely to have resulted from the random sampling process. Either of two conditions are sufficient to insure the validity of the null hypothesis: (a) the sampled populations are identical, (b) the sampled populations are both symmetrical and are symmetrical about a common point. For populations which are identical or symmetrical about a common point, medians, as well as means, are equal. And if the two populations are symmetrical, but symmetrical about different points, or if they have identical forms, but different locations, then medians, as well as means, differ. Thus, if all assumptions are met, the Wilcoxon test is both a test for equality of medians and a test for equality of means.

c. Assumptions. See 1, Fisher's Method of Randomization: Matched Pairs, substituting "mean and median" for "mean".

d. Treatment of Ties. If there are an even number, x, of zero difference scores, consider them to "occupy" the x lowest ranks, give each of them the midrank, and assign half of them a plus sign, half a minus sign in the obtained sample. Thus, if there are x zero differences, each receives the rank $\frac{\sum_{i=1}^{x} i}{x}$ and half of these identical ranks are given a plus, half a minus. If there are an odd number of zero differences, the odd one may be discarded and n reduced by one. Or, all x+1 zero differences may be given the midrank $\frac{\sum_{i=1}^{x+1} i}{x+1}$ and $\frac{x}{2}+1$ of these may be given the algebraic sign least conducive to rejection of the null hypothesis, the remainder receiving the opposite sign.

If nonzero differences are tied in absolute magnitude, the members of each tied group should be given the midrank of the group, i. e., the average rank the members of the group would have if not tied but differing infinitesimally in magnitude. The midrank of each tied member is then given the algebraic sign of that member. An error, which is usually small, is introduced by the occurrence of ties and their treatment in this manner. For example, consider the following set of signed ranks: 1, 2, 3, -4, 5, 6, 7. An equal or smaller negative rank sum can be obtained in six ways and the significance level for the corresponding one-tailed test is $6/2^7$ or .047. However, if the first two ranks are tied, the set becomes 1 1/2, 1 1/2, 3, -4, 5, 6, 7 and there are only five ways of obtaining an equal or smaller rank sum (because 3 and 1 1/2 sum to 4 1/2 while 3 and 1 sum to 4, the value not to be exceeded). The significance level is therefore $5/2^7$ or .039.

The above treatment minimizes error in the long run. To insure that zero or tied differences do not spuriously cause rejection in a specific case, arbitrarily assign the tied-for ranks to each set of tied difference scores (including zero), then give each of the resulting ranks that algebraic sign which is least conducive to rejection of the null hypothesis.

It has sometimes been recommended that all zero differences be discarded and n be reduced accordingly. The reason usually given is that power is greatest if zero differences are treated this way. However, the "increase" in power is quite deceptive since the increase in the probability of rejecting a false null hypothesis is paralleled by an increase in the probability of rejecting a true one. The latter increase raises the actual value of α while its nominal value remains the same. The increase in power is therefore a spurious one which cannot be regarded as an advantage. See "Treatment of Ties" of the Sign test.

e. Efficiency. Asymptotic relative efficiency, compared with Student's t-test when both tests are applied to populations meeting all of the assumptions of the t-test, is $3/\pi$ or .955. The corresponding efficiency for finite samples increases with decreasing sample size, becoming as high as .995 in certain cases. See 3, Test for Location of the Median.

f. Application. Let the following table represent data collected in the application of treatments to pairs of rats from a common population, the pairing having been done on the basis of weight. The null hypothesis is that for each weight category the two treatments have effects which are either identically distributed or are symmetrically distributed about the same median. The alternative hypothesis is that in one or more weight categories the two treatment effects do not have common medians and means.

Treatment **A**	42	37	63	27	46	49	54	39	46	101
Treatment B	42	37	59	34	38	40	43	25	32	33
A-B Difference	0	0	4	-7	8	9	11	14	14	68
Magnitude ranks	1 1/2	1 1/2	3	4	5	6	7	8 1/2	8 1/2	10
Signed ranks	1 1/2	-1 1/2	3	-4	5	6	7	8 1/2	8 1/2	10

The sum of the negatively signed ranks is -5 1/2. A negative sum that small or smaller can be obtained in the following ways:

1 1/2	1 1/2	3	4	-5	6	7	8 1/2	8 1/2	10
1 1/2	1 1/2	3	-4	5	6	7	8 1/2	8 1/2	10
1 1/2	-1 1/2	3	-4	5	6	7	8 1/2	8 1/2	10
-1 1/2	1 1/2	3	-4	5	6	7	8 1/2	8 1/2	10
1 1/2	1 1/2	-3	4	5	6	7	8 1/2	8 1/2	10
1 1/2	-1 1/2	-3	4	5	6	7	8 1/2	8 1/2	10
-1 1/2	1 1/2	-3	4	5	6	7	8 1/2	8 1/2	10
1 1/2	-1 1/2	3	4	5	6	7	8 1/2	8 1/2	10
-1 1/2	-1 1/2	3	4	5	6	7	8 1/2	8 1/2	10
-1 1/2	1 1/2	3	4	5	6	7	8 1/2	8 1/2	10
1 1/2	1 1/2	3	4	5	6	7	8 1/2	8 1/2	10

Thus 11 of the 1024 possible assignments of algebraic sign to the ranks shown above lead to a negative sum as small or smaller than that derived from the obtained sample. The significance level for a one-tailed test of the hypothesis that treatment A produces the same or less "location" effect than treatment B is therefore 11/1024 or slightly greater than .01. For a two-tailed test, there would be 22 assignments giving a sum with absolute value as small as that obtained, and the significance level would be 22/1024 or approximately .02. In practice, significance levels would have been obtained from one of the many tables available and the above enumerations would have been unnecessary. One need only find the sum of the positively signed ranks and the sum of the negatively signed ranks for the obtained sample. The smaller of these two sums in absolute magnitude is referred to prepared tables.

g. Discussion. In analogy with the treatment of Fisher's test, when the Wilcoxon test is used as a test for location it has been assumed that the two sampled populations have either the same form or forms each of which is symmetrical. This means that "treatment", if it produces any effect at all, merely causes a translation or slippage of one distribution relative to the other along the x-axis. Such uncomplicated treatment effects are, in fact, seldom encountered since factors affecting the location of a distribution tend also to affect its dispersion and form. It is reasonable enough to consider that the populations have either identical or symmetrical forms if the null hypothesis is true because a true null hypothesis implies that one of these conditions exists (and implies further that they have identical location parameters). A false null hypothesis does not imply it. Since the assumption is an unrealistic one, it is of interest to examine the likelihood that its failure to be met will cause false acceptance of the alternative hypothesis that the populations differ in location.

If the assumption is dropped, then, when the null hypothesis is false, the true situation may be described by one of a number of alternative hypotheses: (a) the two populations differ in all location parameters and have symmetrical or identical forms, (b) the two populations differ in all location parameters and do not have symmetrical or identical forms, (c) the two populations differ in certain location parameters but not others and do not have symmetrical or identical forms, (d) the two populations have identical location parameters and do not have symmetrical or identical forms. If either (a) or (b) is true the experimenter does not err in accepting the alternative hypothesis that the two populations differ in location. If (d) were true it would mean that two populations in each of which mean and median differed (because the populations are not symmetrical) had equal means and equal medians but different, asymmetrical forms. This requires the unlikely coincidence that two curves with different contours either cross or touch at each of two specified points. The probability for (d) is therefore obviously very small. For (c) however it is required only that different curves, at least one of which is asymmetrical, cross or touch at one of certain specified points. Thus the two populations may have equal means but unequal medians or the reverse. Case (c), therefore, is not at all improbable, and it raises the question, "To which location parameter is the test most sensitive?"

Fisher's test took the mean difference as its test statistic and, in effect, took extreme mean differences as its rejection region. The mean difference is the same as the difference between sample means. There is therefore a direct relationship between the test statistic and the difference between populations means. Fisher's test, therefore, would be expected to be most sensitive to differences between means.

The situation is not nearly so clear cut in the case of the Wilcoxon test. Here the test statistic is neither the difference between means nor the difference between medians, nor does its rejection region consist of such measures. In Fisher's test the average difference score is also the difference between sample means, but in Wilcoxon's test the average signed rank, which is, in effect, the test statistic, does not correspond to any statistic indicating difference in a standard location parameter.

To pursue the question further, if no assumptions other than continuity, randomness and independence were made, Fisher's test would still appear to be a reasonable test for differences in means. The Sign test, which ignores difference-score magnitudes and uses only their direction, i.e., algebraic sign, is obviously the appropriate analogous test for difference in population medians. But Fisher's test, the Wilcoxon test and the Sign test all use the signs of difference scores, differing primarily in the weight which the signs are given prior to summing. For the Sign test the weight is always 1, for the Wilcoxon test it is the rank of the difference-score's absolute magnitude, and for Fisher's test it is the absolute magnitude itself. The Wilcoxon test therefore is intermediate between a test sensitive only to differences in medians and a test sensitive primarily to differences in means. Under the limited assumptions listed above, therefore, the Wilcoxon test should be considered sensitive to both differences in medians and differences in means. Without the asṣumption of symmetrical or identical forms, therefore, it would be futile to attempt to specify which location parameters differ when the null hypothesis is rejected.

Both Fisher's and Wilcoxon's test test the null hypothesis that for every matched pair the observations come from identical populations or populations symmetrically distributed about a common point. It is <u>not</u> assumed that the members of every matched pair are sampled from the <u>same</u> two populations. There may, in fact, be as many pairs of populations as there are difference scores. However, if each pair of units be regarded as equally "important", i.e., to be given equal, a priori weight in determining whether to reject or not, another assumption is required. Under the conditions stated, in order to obtain optimal power it must be assumed that each difference score is as likely to have been obtained from one matched pair of units as from another. This, in turn, means that whatever the variation among the various A-populations or among the n different B-populations, the n difference scores came from identical difference-score populations.

This assumption is analogous to that of homoscedasticity. Without the assumption, if for every matched pair the A and B populations are identical, the pairs whose AB populations have greatest variance are the pairs most likely to have difference scores of large magnitude. These particular pairs will therefore exert greater influence upon the outcome of the test than will those whose AB populations have relatively small variance. When the null hypothesis is false, large difference-score magnitudes resulting from real treatment effects may tend to be cancelled out by large difference-score magnitudes resulting from large population variances and having, by chance, the opposite sign. The power of the test is therefore affected adversely when the assumption is not met.

It has sometimes been claimed that so long as the members of each pair were obtained under matched conditions, the basis for matching may vary from pair to pair. It is clear that such a procedure is quite likely to result in unequal population variances for the various A populations as well as for the B populations and thus, probably, for the population of AB differences. Therefore the power of the test is likely to be altered in such a way that the matching criteria will influence the outcome of the test and the influence of certain of the criteria will be greater than that of others. Furthermore, a certain ambiguity arises when the null hypothesis is rejected because it is not clear what alternative hypothesis is to be embraced. A sample of variously matched scores can only be regarded as representing a multivariate, or at least "multiconditional" population. Therefore, it is this population to which statistical inference must be extended, and conclusions must lack a certain specificity.

To summarize, it is true that the mathematical basis of the Wilcoxon test does not require the assumption that all paired scores were matched on the basis of the same criterion. However, unless such a procedure is followed, the test is likely to be biassed in the sense that certain pairs will yield difference-scores with greater variance, and therefore be given greater influence over the tests outcome, than others, and it is unlikely that the experimenter will know which pairs are so favored. This unknown and unequal influence makes interpretation of the test extremely unclear whether the null hypothesis is rejected or not. And if the null hypothesis is rejected it is not clear what alternative hypothesis to accept because the cause of rejection is uncertain.

h. Tables. Tables can be found in 53, 70, 72, 73, and in some of the sources listed in the introduction. For cases not covered by existing tables, exact probabilities may be calculated by the method of complete enumeration, or approximate probabilities may be obtained from normal tables by treating the rank sum as a normal deviate. Let T be the rank sum for ranks of one sign. Then, if the null hypothesis is true, T comes from a population of rank sums whose mean is $\frac{n(n+1)}{4}$ and whose variance is $\frac{n^2(n+1)}{12}$. As n approaches infinity, the distribution of T approaches the normal distribution. Therefore the approximate probability level for T can be obtained by referring the critical ratio $\dfrac{T - \frac{n\ (n+1)}{4}}{\sqrt{\frac{n^2\ (n+1)}{12}}}$ to normal probability tables. The approximation is reasonably good, when n is large, except at the extreme tails of the normal distribution. Therefore extreme levels of significance, such as the .001, should not be adopted when the normal approximation is used.

i. Sources. 53, 70, 71, 72, 73, 74. See also 5, The Wilcoxon Test: Unmatched Data.

3. Test for Location of the Median

a. Rationale. Let n observations be taken from a continuous, symmetrically distributed population and let the population median be subtracted from each observation. Then the difference-scores constitute a sample of size n from a continuously distributed population symmetrical about a median of zero. Therefore each of the n difference-scores was as likely, before sampling, to be positive as to be negative. And since the populations are continuous, zero differences are not to be expected. Now, rank the difference scores in order of absolute magnitude and give each such rank the algebraic sign of the difference-score whose magnitude it represents. If the true population median was subtracted from each of the n difference scores, the rank sum for ranks of one algebraic sign will have the same distribution as that tabled for the Wilcoxon matched pairs test. In fact, this test may be regarded as a Wilcoxon test in which the A-population is symmetrical and the B-population is a single value, the median of the A-population.

Actually the n observations need not be taken from the same population. Each observation may be drawn from a different population so long as every sampled population is continuous and symmetrical.

b. Null Hypothesis. Each of the 2^n unique sets of signed ranks, obtainable by arbitrarily assigning algebraic signs to the ranks of the difference-score magnitudes, is equally likely to have resulted from the random sampling process. This will be the case if all assumptions are met and if all sampled populations have the same median.

c. Assumptions. Random and independent observations and no zero differences, or preferably continuously distributed populations. (For reasons see 1, Fisher's Method of Randomization: Matched Pairs.) In addition it is assumed that every sampled population is symmetrically distributed. Therefore, if all assumptions are met the null hypothesis can be false, i.e., plus and minus can be unequally likely signs for a difference score, only because the subtracted, hypothesized median is not the true population median.

d. Treatment of Ties. See 2, The Wilcoxon Test: Matched Pairs.

e. Efficiency. Asymptotic efficiency relative to Student's t when both tests are applied to normally distributed populations is $3/\pi$ or .955 (Pitman quoted in 53). Small sample efficiency for same situation appears to vary between .875 and .995 for $n \leq 15$ (53, 64, 65, 66).

f. Application. Subtract the single hypothesized median from each of the n obtained observations. Apply the Wilcoxon matched-pairs test to the difference scores. If the null hypothesis is rejected, conclude that the hypothesized median is not the true median in all of the populations sampled.

Alternatively, apply the Walsh test (see Discussion) to the difference scores, drawing the same conclusion if the null hypothesis is rejected.

g. Discussion. Walsh (64, 65, 66) has outlined a test which Tukey (53) has shown to be equivalent to the above application of the Wilcoxon test. Walsh assumes populations each of which is continuous and symmetrical and tests the hypothesis that all populations have a common specified median. An observation is drawn from each population and the n observations are then ranked in order of algebraic magnitude. The null hypothesis is rejected if certain order statistics (depending on the tail or tails selected for the rejection region) exceed or are exceeded by the hypothesized median. The order statistics used are the averages of two observations of specified rank. The efficiency of the test is high, being the same as that of the Wilcoxon test, and tables (64, 65, 66) are available for small values of n. Tukey has pointed out that the Wilcoxon test is easier to apply when testing the hypothesis of a common median of specified value, while the Walsh test is easier for setting confidence limits for the median. This follows from the manner in which the Walsh test is applied: the null hypothesis is rejected if the hypothesized median falls above or below a difference score of a certain rank or the average of two difference scores whose ranks are specified. The Wilcoxon test, on the other hand, establishes confidence limits by a trial and error method (74). See 53 for exact Walsh method.

h. Tables. Tables listed under 2, The Wilcoxon Test: Matched Pairs, are appropriate. Also 64, 65, and 66 give tables specifically designed for this application and particularly appropriate for setting confidence limits.

i. Sources. 53, 64, 65, 66.

4. Fisher's Method of Randomization: Unmatched Data

a. Rationale. If two samples, of sizes m and n, are random samples from the same population, they may be regarded as a single sample of size m+n which has been drawn from the parent population and then divided on some random, i.e. chance, basis into two subsamples of sizes m and n. If the observations are not matched or paired in any way and if no observations have the same value, there are $\binom{m+n}{n}$ different ways such a "split" could be obtained, and each of these ways is equally likely.

Now suppose that for each "way" some statistic, say the mean, is calculated for each of the two subsamples and the difference $\overline{X}_A - \overline{X}_B$ obtained, the subscripts A and B being arbitrary labels to identify the two subsamples. If N of these $\overline{X}_A - \overline{X}_B$ differences equal or exceed the $\overline{X}_A - \overline{X}_B$ difference for the actually obtained samples, then the chance probability of the actually obtained $\overline{X}_A - \overline{X}_B$ difference or one more extreme among the differences calculated for the $\binom{m+n}{n}$ "splits" is $N/\binom{m+n}{n}$.

If the two original samples were actually obtained under two different treatments, then if the treatments have equal effects, the samples are, in effect, samples from the same population. Thus the hypothesis of identical treatment effects can be tested at the α level of significance by rejecting the hypothesis if $N/\binom{m+n}{n} \leq \alpha$.

b. <u>Null Hypothesis</u>. Each of the $\binom{m+n}{n}$ different pairs of "samples" obtainable by dividing the total of m+n observations into two sets, one containing m observations, the other n observations, is equall y likely to have been obtained in the experiment. A sufficient condition for the validity of the null hypothesis is that the two sampled populations are identically distributed. This will be the case if treatments do not differ in their measured effects on individuals and if individuals are assigned randomly to treatments. By taking as the rejection region the N pairs of sets with the N greatest mean differences, the method of randomization tests the null hypothesis that populations are identical and is "most sensitive" to the alternative hypothesis that the populations have different means.

c. <u>Assumptions.</u> Bias in the sampling process or possible influence of one sampled observation upon another may cause some of the $\binom{m+n}{n}$ pairs of rearranged samples to be more likely than others to have been the pair actually drawn. And this may be the case even though all observations in both samples are drawn from the same population. Therefore, in order to confine the cause of unequal probability to failure of the null hypothesis, it is necessary to assume that sampling is <u>random</u> and observations are <u>independent.</u>

If any of the m+n observations have the same value there will be less than $\binom{m+n}{n}$ distinguishable rearrangements of observations into samples of sizes m and n. Thus the sample space for the test statistic will be smaller than that represented by the denominator of the probability fraction. In order to "eliminate" such an eventuality, it is assumed that there are no tied observations. This assumption is sometimes expressed in its mathematically equivalent form: populations are continuously distributed.

If the two sampled populations do not have identical forms, the $\binom{m+n}{n}$ pairs of hypothetical samples may be, and probably are, unequally probable even though the two populations have the same mean. For example, if the two populations are normally distributed with the same mean but different variances, the "splits" which give the more extreme observations to the "sample" from the population with the greater variance are more probable than are the "splits" which do the opposite. Furthermore, if the two populations have both unequal means and different forms, the inequality of means may bias the probability in one direction and the dissimilarity of form may bias it in the opposite direction. Thus the two causes of unequal probability may tend to balance one another. It is extremely unlikely that this balance would be complete, leaving each of the $\binom{m+n}{n}$ pairs of samples equally probable. However the power of the test would be adversely affected. In order, therefore to confine the cause of failure of the null hypothesis to inequality of population means, the alternative hypothesis, it is assumed that, whatever their location, the two sampled populations have identical forms.

Since the last named assumption is a fairly unrealistic one, the experimenter may prefer to substitute the more reasonable assumption that if population means are equal their forms are identical. Thus any dissimilarity of form must be accompanied by an inequality of means, and the null hypothesis can be false only when means differ. When the null hypothesis is false, then the alternative hypothesis of unequal means must be true. However, the power of the test to detect the validity of the alternative hypothesis may be much smaller than would be the case if identical forms could be legitimately assumed.

d. Treatment of Ties. If a small proportion of the observations are tied it may be reasonable to suppose that the ties are attributable to the discreteness of the measuring instrument rather than lack of continuity in the distribution of the thing measured. Therefore, treat each tied observation as though it were unique in determining N, and use $\binom{m+n}{n}$ unaltered, as the denominator of the probability fraction.

e. Efficiency. High efficiency for this test is suggested by the high efficiency of the Wilcoxon test which is a modification of it. See 5, The Wilcoxon Test: Unmatched Data.

f. Application. To modify an example given by Fisher, suppose that the height, in centimeters, has been measured for 8 Englishmen and 7 Frenchmen, and that it is desired to test the hypothesis that Englishmen and Frenchmen have the same average height.

									$\bar{X}$
Englishmen:	188	182	178	177	176	174	173	170	177.25
Frenchmen:	172	171	169	165	164	162	160		166.14

$$\bar{X}_E - \bar{X}_F = 11.11$$

There are $\binom{15}{7}$ or 6435 different ways of reassigning the height measurements so as to give eight of them to Englishmen, seven to Frenchmen. In only four of them will the Englishmen's mean exceed the Frenchmen's mean by a value as great as that obtained in the actual samples:

									$\bar{X}$
Englishmen:	188	182	178	177	176	174	173	172	177.50
Frenchmen:	171	170	169	165	164	162	160		165.86

$$\bar{X}_E - \bar{X}_F = 11.64$$

Englishmen:	188	182	178	177	176	174	173	171	177.375
Frenchmen:	172	170	169	165	164	162	160		166.00

$$\bar{X}_E - \bar{X}_F = 11.375$$

									$\overline{X}$
Englishmen:	188,	182,	178,	177,	176,	174,	173,	170	177.25
Frenchmen:	172,	171,	169,	165,	164,	162,	160		166.14

$\overline{X}_E - \overline{X}_F =$ 11.11

Englishmen:	188,	182,	178,	177,	176,	174,	172,	171	177.25
Frenchmen:	173,	170,	169,	165,	164,	162,	160		166.14

$\overline{X}_E - \overline{X}_F =$ 11.11

Thus the significance level for a one tailed test of the hypothesis that the average Frenchman is as tall or taller than the average Englishman is 4/6435 and the hypothesis could be rejected at an extreme level of significance. Since the hypothesis is that Englishmen and Frenchmen have equal average heights, there are, in addition to the four ways, in which so great a mean difference could be obtained in favor of the Englishmen, the following four ways in which so extreme a mean difference can be found in favor of the Frenchmen.

									$\overline{X}$
Englishmen:	172,	171,	170,	169,	165,	164,	162,	160	166.625
Frenchmen:	188,	182,	178,	177,	176,	174,	173		178.286

$\overline{X}_E - \overline{X}_F =$ -11.661

Englishmen:	173,	171,	170,	169,	165,	164,	162,	160	166.750
Frenchmen:	188,	182,	178,	177,	176,	174,	172		178.143

$\overline{X}_E - \overline{X}_F =$ -11.393

Englishmen:	173,	172,	170,	169,	165,	164,	162,	160	166.875
Frenchmen:	188,	182,	178,	177,	176,	174,	171		178.000

$\overline{X}_E - \overline{X}_F =$ -11.125

Englishmen:	174,	171,	170,	169,	165,	164,	162,	160	166.875
Frenchmen:	188,	182,	178,	177,	176,	173,	172		178.000

$\overline{X}_E - \overline{X}_F =$ -11.125

Thus for a two-tailed test of the null hypothesis that there is no difference between the average heights of Englishmen and Frenchmen, the significance level is 8/6435.

It happened in the above example that the number of mean differences as great as the absolute value of the obtained mean difference is the same for positive as for negative mean differences. This is certain to be the case only when m = n. When the two samples are of unequal size, the significance level for a two-tailed test is not necessarily twice that for a one-tailed test, because symmetry no longer obtains.

g. Discussion. Many of the points requiring discussion are highly analogous to those discussed under 1, Fisher's Method of Randomization: Matched Pairs ; therefore, the arguments will not be repeated here.

Obviously the Method of Randomization is not restricted to testing for differences between means. The significance of a variety of "difference" statistics calculated from two samples can be tested by "calculating" the statistic for each of the $\binom{m+n}{n}$ splits and taking as the rejection region those N splits for which the calculated statistic has the N most extreme values, the significance level, α, being $N/\binom{m+n}{n}$. The "most extreme" values are of course those most suggestive that the alternative hypothesis, rather than the null hypothesis, is true. The alternative hypothesis states, in effect, that the population statistic corresponding to the statistic calculated from the obtained samples is not zero. However, unless the sample statistic can be expected to "represent" well its population counterpart, the power of the test may be very small. For example, the method could not be expected to provide a powerful test for a difference in population ranges.

Pitman (41, 42, 43) has elaborated upon the method of testing for a difference between population means and has applied the Method of Randomization to testing the significance of a correlation coefficient (See next chapter) and to testing the effect of treatments in an analogy of analysis of variance. The latter problem has also been investigated by Welsh (67, 68).

To test for treatment effects in analogy with analysis of variance, Pitman takes m batches (letters) of n individuals each of which is subjected to a different one of n treatments (numbers), the assignment of individuals to treatments being random. The scores of the individuals can be represented as follows:

$$\begin{array}{cccc} a_1, & a_2, & \ldots. & a_n \\ b_1, & b_2, & \ldots. & b_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ m_1, & m_2 & \ldots. & m_n \end{array}$$

If there is no treatment effect, the n scores in each row are randomly placed in the n "treatment" columns. There are n! ways in which the observations in a row can be permuted and since there are m rows, there are $(n!)^m$ tables which can be obtained by permuting the observations within rows. However, some of these tables differ only in the permutation of identical columns. This can be prevented by permitting permutation of observations in all but the last row. Therefore, there are $(n!)^{m-1}$ ways in which the mn observations can be assigned so that each column contains one observation from each batch and so that no two assignments are identical except for the location of columns with respect to each other. Pitman calculates the equivalent of the F ratio for each such assignment and rejects the hypothesis of no treatment effect at the significance level $\alpha = N/(n!)^{m-1}$ if the F ratio for the actually obtained sample lies among the N most extreme of these.

h. Tables. None. Probabilities must be calculated for each specific case.

i. Sources. 4, 7, 16, 26, 27, 28, 34, 39, 40, 41, 42, 43, 48, 67, 68, 75. See also 17 and 38 under 1, Fisher's Method of Randomization: Matched Pairs.

5. The Wilcoxon Test: Unmatched Data

a. *Rationale.* Wilcoxon has modified Fisher's method by replacing the obtained scores with their ranks. The test statistic, which in Fisher's method was the difference in sample means, is, in Wilcoxon's test, the rank sum for the smaller sample, or when samples are of equal size, the smaller of the two sample rank sums. The Wilcoxon modification has advantages similar to those discussed in "Rationale" of 2, The Wilcoxon Test: Matched Pairs; the test is not a conditional one since the sample space for the test statistic is the same for every pair of samples, the test is less sensitive to extreme observations, and the probabilities can be tabled.

b. *Null Hypothesis.* Each of the $\binom{m+n}{n}$ pairs of "artificial" samples obtainable by arbitrarily assigning m observations to one treatment, n to the other, is equally likely to have been drawn as a pair of true samples. If all assumptions are met, a sufficient condition for the validity of the null hypothesis is that the two samples come from identical populations. This will be the case if the two treatments do not differ in any measured respect.

c. *Assumptions.* It is assumed that sampling is *random*, observations are *independent*, *no observations* are *tied* or populations are *continuously distributed*, populations have *identical forms* (or at least have identical forms if population means or medians are equal). For reasons, see 4, Fisher's Method of Randomization: Unmatched Data.

d. *Treatment of Ties.* If ties are due only to imprecision of measurement, i.e., if the thing measured is continuously distributed, then ties are a problem only when members of a tied group lie in both obtained samples. When all the observations have a given tied value lie in one sample, they may be arbitrarily assigned the ranks they would have if distinguishable. If observations in *both* samples have the same value, one technique is to assign tied observations the tied-for ranks least conducive to rejection of the null hypothesis. Another technique is *randomly* to assign to the members of the tied group the ranks they would have if distinguishable. This preserves the mathematical integrity of the test, but forceably and artificially introduces an element of chance which must, in general, reduce the power of the test.

The most frequently recommended technique is to give each of the members of a tied group the midrank of the group, i.e., the average of the ranks the tied members would have if their values were distinguishable. The result is that the set of ranks obtained in this manner and rank sums obtained by applying Fisher's method to them are not the same as the set of ranks and rank sums used (by applying Fisher's method to the m+n different integers from 1 to m+n) to calculate the probability tables. The tables therefore are inaccurate in such cases, giving not the true probability but rather the probability of the average value taken by the test statistic when ties are broken in all possible ways. (If all the observations having the same value lie in the same sample, all ways of breaking ties result in the same value for the test statistic and the tables are fully applicable if discontinuity is due only to imprecision of measurement.)

When midranks are used the rank sum may not be an integer. The tabled rank sums, however, are integers. Therefore, it is suggested that when the obtained rank sum is not an integer it should be raised or lowered one half unit so as to assume whichever integral value is least conducive to rejection of the null hypothesis. This procedure results in a slightly more conservative test.

In many cases the effect of using midranks is very much the same as if tied observations were assigned consecutive ranks with the ranks carefully apportioned so as to "balance" the apportionment between the two samples. For example, suppose ten observations are tied for 21st to 30th place in rank and two of the observations are in sample A, the remainder in sample B. In "balancing" one might assign the ranks 24 and 27 to the two observations in sample A because they separate the ranks 21 to 30 into nearly equal parts, or 21 and 30, 25 and 26 or any other assignment resulting in a "symmetrical" pattern might be picked. The result of course is that in every case the average of these ranks, for each sample, is the midrank, 25 1/2. Therefore, when "symmetrical rank patterns" can be obtained without resorting to nonintegral ranks, the use of midranks is equivalent to assigning to each member of a tied group a different one of the ranks for which the group is tied and doing so in such a way that each sample gets its "fair share" of rank magnitude. If the rank sum is an integer the tables give the exact probability under the assumption that one of the possible "equitable apportionments" is the correct one. In the long run the average difference between this probability and the true probability will tend to be zero; however, in any specific experiment a discrepancy of zero is quite unlikely. Therefore, for the particular experiment under test the probability of false rejection of the null hypothesis may be greater or less than that indicated by the tables. Regardless of whether or not "symmetry" can be obtained with integers, the limits of "tie-error" can easily be found. This is accomplished by assigning tied observations the "tied-for" ranks least conducive to rejection of the null hypothesis, performing the conservative test, then assigning them the ranks most conducive to rejection and performing the radical test, thereby obtaining bounds for the influence of ties on probability levels. This procedure has been recommended by van der Vaart (58) who observes that if the chosen significance level does not lie between these bounds there is no problem and if it does, there is no solution. He adds that precisely the same dilemma arises when ties occur in the application of Student's t-test although "this fact has always passed unnoticed."

When samples are so large that tables are inapplicable the normal approximation is generally used. The difference between the obtained and the expected rank sum is divided by the standard deviation of the rank sum, and the resulting critical ratio is treated as a normal deviate with zero mean and unit variance and referred to normal probability tables. When ties are given the midrank, the presence of ties has no effect upon the expected rank sum, but does affect the variance, causing it to be smaller than would be the case if there were no ties. There is a formula, however, which takes account of ties in calculating variance and therefore "corrects" for ties when used in calculating the critical ratio. This formula requires that the Mann-Whitney form of the Wilcoxon test be used (See 6, The Mann Whitney Test).

e. Efficiency. The value $3/\pi$ or .955 has been obtained for the asymptotic efficiency of the Wilcoxon test relative to Student's t-test when both tests are applied to samples from normally distributed populations with homogeneous variances. This value has been obtained by a number of authors (9, 11, 37, 50, 59, 62), Pitman (not referenced) apparently having been the first, and is true of both one-sided and two-sided tests under several different definitions of asymptotic efficiency. Hodges and Lehmann (23) have shown that the asymptotic relative efficiency of the two-sample Wilcoxon test relative to Student's t cannot fall below .864 when both are used as tests against shift of a continuous, but otherwise unspecified, distribution function. (The comparison is less favorable to the Wilcoxon test when shift is accompanied by "contaminations"). They conclude that to the extent that the concept of asymptotic relative efficiency "adequately represents what happens for the sample sizes and alternatives arising in practice, this result shows that use of the Wilcoxon test instead of Student's t-test can never entail a serious loss of efficiency for testing against shift. (On the other hand ..... the Wilcoxon test may be infinitely more efficient than the t-test.)" In fact Pitman is quoted (23, 47) as having found an A. R. E. of 1 for Wilcoxon's relative to Student's test when both were applied to uniform distributions. Pitman (23, 47) and Pitman and Noether (7) are quoted as having found the A. R. E. of Wilcoxon's relative to Student's test to be considerably greater than 1 when the two tests were applied to certain types of distributions. Similar results have also been found for small samples. Student's test has been found to have power inferior to that of the Wilcoxon test for testing samples of 4 and 6 observations from certain uniform distributions (63) and for testing samples of 5 and 5 from certain distributions differing in peakedness (Whitney quoted in 1).

When both tests are applied to samples from normal populations with homogeneous variances, Student's test has invariably been found to have power as great or greater than Wilcoxon's; however, the difference in efficiency has, with one exception, always been very slight (9, 23, 52, 59, 60, 62). The exception (9) has been criticized (23) as attributable to a procedural artifact.

The evidence therefore supports the conclusion that Student's t-test is statistically more efficient than Wilcoxon's test when the assumptions of the t-test have been completely met, but that the superiority of the t-test is slight, amounting to less than 5%. When Student's assumptions have not been fully met, either test may be the more powerful, depending upon a number of factors. However, if it is known that the populations have identical, continuous forms when their location parameters are equal (i.e, that if treatments have different effects, these include effects upon means and medians), or if the experimenter is interested in detecting <u>any</u> discrepancy between continuously distributed populations (i.e., any type of treatment effect), then the Wilcoxon test is preferable. Rejection of the null hypothesis can occur <u>only</u> because of the existence of the effect in which the experimenter is interested or because of chance with probability of exactly α. If Student's test were used in the same cases, rejection could occur because of (a) the effect whose detection is desired, (b) nonnormality, (c) chance, with probability other than α (and unknown) unless the populations are known to be normal.

The Wilcoxon test is one of the most powerful distribution-free tests. Tests designed by Terry and van der Waerden, and discussed in the Introduction and in (7) are slightly more efficient, in the statistical sense, for certain test situations. However, they lack the Wilcoxon test's conceptual simplicity and ease of application. In several investigations of the power of distribution-free tests with respect to each other, the Wilcoxon test has invariably been found to be most powerful or among the most powerful (See Table II in Introduction).

Mann and Whitney (35) showed that the Wilcoxon test is consistent "with respect to the class of alternatives f (x) > g (x) for every x", i.e., is consistent if the alternative to the null hypothesis of identical populations is that the cumulative distribution of one population lies entirely above, i.e. does not cross, that of the other. Van Dantzig (6) and Lehmann (32) have pointed out that Mann and Whitney's proof actually is more general. It proves the test consistent if, when the null hypothesis is false, the probability that a random observation from one population exceeds one from the other population differs from 1/2 (for a two-tailed test or, for a one-tailed test, differs from 1/2 in a specified direction) (30). The above results require that the ratio m/n remain constant as $n \to \infty$. Putter (44) has shown that, under the same conditions, if the populations are discontinuous and $\Pr(x > y) + 1/2 \Pr(x = y) > 1/2$ the test will be consistent if ties are randomized, i.e., if ties in each group of tied observations are randomly assigned the tied-for ranks.

Lehmann (32) has proved that the Wilcoxon test is unbiassed when it is used as a one-tailed test, more specifically it is unbiassed for the class of alternatives F (x) > G (x) for every x. Van der Vaart (55, 59) has shown that the two-tailed Wilcoxon test may be, but is not necessarily, biassed. The likelihood of such bias appears to be greater when samples are of unequal size and when populations are skewed.

Mann and Whitney (35) showed that their mathematically equivalent test statistic is asymptotically normally distributed under the null hypothesis if m and n approach infinity in any arbitrary manner. Lehmann (32) has found that it is also asymptotically normally distributed when the populations differ provided that the ratio m/n remains constant as m and n approach infinity. Stoker (49) states that Lehmann's proof also applies when populations are discontinuous.

Asymptotic normality has also been proven by Haldane and Smith (20).

f. Application. Suppose that gain in weight has been measured under two different diets with the following results for six individuals subjected to Diet A and seven persons given Diet B.

Diet A		Diet B	
Weight Gain	Rank	Weight Gain	Rank
-14	1	-3	5
-12	2 1/2	5	8
-12	2 1/2	7	9
-10	4	8	10
- 2	6	9	11
2	7	15	12
		24	13
Sum	23	Sum	68

There are $\binom{6+7}{6}$ or 1716 ways of redistributing the scores into samples of sizes 6 and 7. Of these, there are only four ways in which Diet A could obtain a rank sum equal to or smaller than the obtained rank sum of 23. They are as follows (only the ranks being shown):

1, 2 1/2, 2 1/2, 4, 5, 6	$\Sigma = 21$
1, 2 1/2, 2 1/2, 4, 5, 7	$\Sigma = 22$
1, 2 1/2, 2 1/2, 4, 5, 8	$\Sigma = 23$
1, 2 1/2, 2 1/2, 4, 6, 7	$\Sigma = 23$

The significance level for a one-tailed test of the hypothesis that Diet A causes the same or more weight gain than Diet B, therefore, is 4/1716 or about .0023.

Since the samples are of unequal size, a two-tailed test raises the question of which rank sums to consider as extreme in the opposite direction. Obviously they cannot be those totaling to 68 or more for Diet A, because that number was obtained for Diet B as the sum of seven ranks, while for Diet A only six ranks can be summed. The solution proposed by White (69) is to rerank the observations, this time ranking the largest observation 1, the next largest, 2, etc.; then the number of ways of redistributing scores which cause Diet A to have a rank sum of 23 or smaller are those whose rank sums are as extreme or more extreme in the "opposite direction." There are, in fact, four such ways and the probability level for a two-tailed test is therefore 8/1716. However, the reranking need not actually be performed because the test statistic is symmetrically distributed and the probability level for a two-tailed test is simply twice that for a one-tailed test.

In practice, of course, probabilities would generally not be obtained by applying the method of randomization, but would be obtained from tables. In that case, only the rank sums need be obtained. The use of tables varies considerably, however, from one table to another, and the particulars of application will not be described here.

g. Discussion. Various forms of the Wilcoxon test have been published by a variety of authors. Wilcoxon developed the test for the case where samples are of equal size, i.e., $m = n$. White (69) extended the test, and tabled it, to the case of unequal sample sizes. This was also done by van der Reyden (45) at about the same time, but apparently without knowledge of the work of either Wilcoxon or White. A test, conducted differently, but mathematically equivalent to the Wilcoxon test, was developed independently by Festinger (15) and published very soon after Wilcoxon's original article. Festinger took as his test statistic the absolute difference between the average rank for the smaller sample and the average rank for the combined sample of m+n ranks. Since the latter is a constant (equal to $\frac{m+n+1}{2}$) for fixed values of m and n, and since the average rank for the smaller sample is simply its rank sum divided by its size, Festinger's test is mathematically equivalent to White's extension of the Wilcoxon test. Because of the additional computation required to obtain the test statistic, d , Festinger's test is more time consuming than the Wilcoxon test. A Wilcoxon-like test was developed by Haldane and Smith (20, see also 3 and 24) for a specific application. Finally, a modified form of the Wilcoxon test developed by Mann and Whitney (35) has become the most widely used form of the test. It is discussed in the next section. Because of the mathematical relationships existing between the Wilcoxon, White, van der Reyden, Festinger and Mann-Whitney tests, they have common mathematical properties of efficiency, consistency, asymptotic normality, etc.

The Wilcoxon test actually tests whether or not two populations are identical. The test becomes a test for equal means (or substitute "medians") if it can be legitimately assumed either (a) that whatever their locations the populations have identical forms, or (b) that if their means (or substitute "medians") are equal the populations have identical forms, i.e., the populations are identical. The latter assumption is generally far more realistic than the former; however, the test may have les s power if only the latter assumption can be made. See "Assumptions" under Section 4, Fisher's Method of Randomization: Unmatched Data.

Wilcoxon (71, 72, 73, 74) has extended his test to permit a single test of data collected under several, different, non-tested experimental conditions. Under each of k non-tested conditions, n observations are taken under treatment A and n observations under treatment B. Then, except for the last step of determining significance levels, the ordinary Wilcoxon test is performed for each non-tested condition independently. This results in a rank sum, based on n ranks, for treatment A, and one for treatment B, under each of the k non-tested conditions. The sum of the k rank sums is then obtained for each treatment and the smaller of these is referred to a brief, specially prepared table of probabilities. The test is legitimate (as a test for simple treatment effects) provided that when the k non-tested conditions have different effects upon observations, any given condition has the same effect upon observations taken under one treatment as it has upon observations taken under the other. That is to say, there must be no interaction between treatments and non-tested conditions. If this implicit assumption is not met, the power of the test may be adversely affected and when the null hypothesis (that each of the k B-populations has the same form and location as its A-population counterpart) is false, the true alternative hypothesis will be unable to be specified in other than very general terms.

h. Tables. Tables can be found in 45, 69, 70, 72, 73 for the Wilcoxon or rank sum form of the test, in 15 for Festinger's difference-in-average-rank form, and in 1, 35, 46 (and see also 18 and 36) for the Mann-Whitney form of the test. Tables for Wilcoxon's application of his test to data collected under a variety of non-tested conditions are in 71, 72 and 73. Tables can also be found, reproduced, in some of the sources listed in the Introduction.

Several of these tables have been found to contain errors.

Auble's tables have been criticized by Fix and Hodges (18), Festinger's tables by Kruskal and Wallis (30), van der Reyden's tables by Kruskal and Wallis (31), and White's Tables by Fix and Hodges (18) and Kruskal and Wallis(31).

For cases not covered by existing tables, probabilities may be obtained by the method of randomization, or the rank sum may be treated as a normal deviate and approximate probabilities may be obtained by referring a critical ratio to normal tables. Let T be the rank sum for the sample with m observations. Then, if the null hypothesis is true, T comes from a population of rank sums whose mean $\bar{T}$ is $m\left(\frac{m+n+1}{2}\right)$ and whose variance σ_T^2 is $m\,n\left(\frac{m+n+1}{12}\right)$. As m and n increase, the distribution of T approaches the normal distribution. Therefore, the approximate probability level for T can be obtained by referring the critical ratio $\dfrac{T - m\left(\frac{m+n+1}{2}\right)}{\sqrt{m\,n\left(\frac{m+n+1}{12}\right)}}$ to normal probability tables. The approximation is reasonably good, when m and n are large, except at the extreme tails of the normal distribution. Therefore extreme levels of significance, such as the .001, should not be adopted when the normal approximation is used.

If T is the rank sum for the sample with m observations when the smallest rank is assigned to the smallest observation, and T' is the rank sum for the same sample when the smallest rank is assigned to the largest observation, then $T' = m(m + n + 1) - T$. This is easily seen: If r is the rank of one of the m observations in the first case and r' is the corresponding rank in the second case, then $r' = m + n - (r - 1) = m + n + 1 - r$. And since $T' = \Sigma_1^m r'$, then $T' = \Sigma_1^m (m + n + 1) - r = m\,(m + n + 1) - \Sigma_1^m r = m\,(m + n + 1) - T$. This formula saves the labor of reranking when tables, such as White's, require the smaller of the two T values.

i. Sources: 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18, 19, 20, 21, 22, 23, 24, 25, 29, 30, 31, 32, 33, 35, 36, 37, 44, 45, 46, 47, 49, 50, 51, 52, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 69, 70, 71, 72, 73, 74.

6. The Mann Whitney Test

a. Rationale. Let a sample of n observations, designated as Xs, and a sample of m observations, identified as Ys, be taken from the same continuously distributed population. Now arrange the m + n observations in order of increasing size irrespective of sample. Then replace each ordered observation with an X or a Y depending on the sample from which it originally came. The result will be a pattern of n X's and m Y's intermixed.

If these m + n units were all different, there would be (m + n)! distinguishable patterns. However, for each actually distinguishable pattern there are n! permutations of Xs with each other which do not change the pattern, and for each of these permutations there are m! permutations of Y's with each other which do not change the pattern. Therefore, there are $\frac{(m+n)!}{m!\,n!}$ or $\binom{m+n}{m}$ distinguishable patterns of n Xs and m Ys. If the two samples are drawn from the same population each of these patterns is equally likely. However, if they come from different populations, the patterns should be unequally likely, and if the populations differ in location only, one would anticipate patterns in which Xs tended to cluster at one end, Ys at the other.

The test statistic, U, therefore is the number of times a Y precedes an X. Thus, U is the number of Ys preceding the smallest X, plus the number of Ys preceding the next smallest X (and therefore including all of the Ys counted in the first case), etc., until the number of Ys preceding each X are counted and summed for all Xs. The probability of U, when the null hypothesis is true, is simply the proportion of the $\binom{m+n}{m}$ possible patterns which result in Us as extreme or more extreme than that obtained.

b. Null Hypothesis. Each of the $\binom{m+n}{m}$ patterns of Xs and Ys, representing their observations arranged in order of increasing algebraic magnitude, is equally likely. A sufficient condition for the validity of the null hypothesis is that the two samples were drawn randomly and independently from identical continuously distributed populations.

c. Assumptions: See 5, The Wilcoxon Test: Unmatched Data.

d. Treatment of Ties. If Xs are tied with Ys, and m and n are small enough for the tables to apply, it is suggested that the

Wilcoxon form of the test be used and that ties be treated as outlined in 5, The Wilcoxon Test: Unmatched Data.

If m and n are large enough to justify using the normal approximation to the distribution of U, a correction for ties can be applied in calculating the critical ratio. See "Tables".

e. Efficiency. See 5, The Wilcoxon Test: Unmatched Data. Efficiency, power, consistency and bias are same as for the Wilcoxon test for unmatched data.

f. Application. Let the observations from the example of application of the Wilcoxon test be arranged in order of increasing magnitude, with the letter in parentheses indicating the sample from which an observation came. The result is -14(A), -12(A), -12(A), -10(A), -3(B), -2(A), 2(A), 5(B), 7(B), 8(B), 9(B), 15(B), 24(B). The number of times a B precedes an A is 2. A value of U as small or smaller than this could be obtained from the following arrangements:

A A A A A A B B B B B B B U = 0

A A A A A B A B B B B B B U = 1

A A A A A B B A B B B B B U = 2

A A A A B A A B B B B B B U = 2

Since there are $\binom{6+7}{6}$ or 1716 possible arrangements, the significance level for a one-tailed test of the hypothesis that the A's either equal or exceed the B's is 4/1716. For a two-tailed test, the mirror images of the four patterns shown above must be considered as causing large U's which are correspondingly "as extreme". These are the patterns in which a B follows an A zero, 1, 2, and 2 times, or, to return to the definition of U, the ways in which a B precedes an A 42, 41, 40 and 40 times. Since there are eight values of U as extreme as that obtained, the values being 0, 1, 2, 2, 40, 40, 41, 42, the significance level for a two-tailed test is 8/1716.

g. Discussion. Let there be n xs and m ys arranged in order of increasing magnitude. Let x_i be the i^{th} x in order of increasing magnitude and the r^{th} measurement, i. e., the r^{th} among the xs and ys combined, in order of increasing magnitude, and let u_i be the number of ys preceding x_i. Finally let T be the Wilcoxon rank sum of the x ranks and let U be the Mann-Whitney statistic, the number of times a y precedes an x. Then r is the Wilcoxon rank of x_i and

$r = i + u_i$. And $T = \Sigma r = \Sigma_{i=1}^{n} (i + u_i) = n(\frac{n+1}{2}) + \Sigma_{i=1}^{n} u_i = n(\frac{n+1}{2}) + U$.

The sum of all ranks is simply the number of ranks times the average rank, or $(m+n)(\frac{m+n+1}{2})$; therefore, T', the rank sum of the y ranks is $(m+n)(\frac{m+n+1}{2}) - T$. So $T' = (m+n)(\frac{m+n+1}{2}) - n(\frac{n+1}{2}) - U$ which reduces to $T' = mn + \frac{m(m+1)}{2} - U$.

Thus the Mann-Whitney test statistic U, for any given values of m and n, differs from the Wilcoxon test statistic, T, only by a constant. Otherwise stated, the two statistics are mathematically equivalent. The formulas relating T to U may be useful in saving labor when tables are in terms of U, since it is generally easier to obtain T than U (which involves an excessive amount of counting). The Mann-Whitney statistic is also related to Kendall's S for rank correlation.

Many of the points discussed in connection with Fisher's Method of Randomization and the Wilcoxon test are also relevant to the Mann-Whitney statistic. They will not be recapitulated; therefore, see the "Discussion" section of the foregoing tests named.

h. Tables. 1, 35, 46 (See also 18 and 36). Tables can also be found, reproduced, in some of the sources listed in the Introduction.

The number of ys which either precede or follow a given x is m, the size of the y sample; and since there are n xs, the number of ys either preceding or following an x is nm. Therefore if U is the number of times a y precedes an x, then mn - U is the number of times a y follows an x. This, however, is also the number of times an x precedes a y. Therefore, the count need be made only once even though most tables list only the smaller of the two values U and U' = mn - U.

When m and n are large and are too large for the exact tables to apply, approximate probabilities may be obtained by referring a critical ratio to normal tables. If there are no ties, the test statistic U comes from a population of U's whose mean is $\overline{U} = \frac{mn}{2}$

and whose variance is $\frac{mn(m+n+1)}{12}$. Ties do not affect the mean, but they decrease the variance (22). Let t_i be the number of tied observations in the i^{th} group of tied observations, and let there be k groups. Then when there are ties the variance becomes

$$\sigma_u^2 = \frac{mn}{12} \left[m+n+1 - \frac{\Sigma_{i=1}^{k} (t_i^3 - t_i)}{(m+n)(m+n-1)} \right].$$

Probabilities may therefore be obtained by referring the critical ratio $\frac{U - \overline{U}}{\sigma_u}$ to normal tables.

i. Sources. See 5, The Wilcoxon Test: Unmatched Data

BIBLIOGRAPHY

T 1. AUBLE, D., Extended tables for the Mann-Whitney statistic. Bulletin Institute Educational Research, Indiana University, 1953. Vol. 1.

2. Baten, W. D. and Trout, G. M., A critical study of the summation-of-difference-in-rank method of determining proficiency in judging dairy products. Biometrics, 1946, 2, 67-69.

3. Bennett, B. M., The power function of the Haldane-Smith test. (Abstract) Annals of Mathematical Statistics, 1952, 23, 476.

4. Camp, B. H., Some recent advances in mathematical statistics, I. Annals of Mathematical Statistics, 1942, 13, 62-73.

5. van Dantzig, D., Asymptotische eigenschappen van Wilcoxon's toets. (English summary) Published lecture notes for material published also in (6).

6. van Dantzig, D., On the consistency and power of Wilcoxon's two sample test. Proceeding Koninklijke Nederlandse Akademie van Wetenschappen (Ser. A), 1951, 54, 1-8.

7. van Dantzig, D. and Hemelrijk, J., Statistical methods based on few assumptions. Bulletin of the International Statistical Institute, 1954, Vol. 34.

8. David, Florence N., A note on Wilcoxon's and allied tests. Biometrika, 1956, 43, 485-488.

9. Dixon, W. J., Power under normality of several nonparametric tests. Annals of Mathematical Statistics, 1954, 25, 610-614.

10. Dwass, M., On the asymptotic normality of certain rank order statistics. Annals of Mathematical Statistics, 1953, 24, 303-306.

11. Dwass, M., The large-sample power of rank order tests in the two-sample problem. Annals of Mathematical Statistics, 1956, 27, 352-374.

12. van Eeden, Constance and Benard, A., A general class of distributionfree tests for symmetry containing the tests of Wilcoxon and Fisher. I, II, and III. Proceedings Koninklijke Nederlandse Akademie van Wetenscahppen (Series A), 1957, 60, 381-391, 392-400, 401-408.

13. Epstein, B., Comparison of some non-parametric tests against normal alternatives with an application to life testing. Journal of the American Statistical Association, 1955, 50, 894-900.

14. Epstein, B., Tables for the distribution of the number of exceedances. Annals of Mathematical Statistics, 1954, 25, 762-768.

T* 15. Festinger, L., The significance of difference between means without reference to the frequency distribution function. Psychometrika, 1946, 11, 97-105.

* 16. FISHER, R. A., "The coefficient of racial likeness" and the future of craniometry. Journal of the Royal Anthropological Institute of Great Britain and Ireland, 1936, 66, 57-63.

* 17. FISHER, R. A., The design of experiments, New York: Hafner, 1953, (Sixth Ed.), 43-47.

T 18. Fix, Evelyn and Hodges, J. L., Significance probabilities of the Wilcoxon test. Annals of Mathematical Statistics, 1955, 26, 301-312.

19. Fraser, D. A. S., Non-parametric theory: scale and location parameters. Canadian Journal of Mathematics, 1954, 6, 46-68.

* 20. Haldane, J. B. S. and Smith, C. A. B., A simple exact test for birth-order effect. Annals of Eugenics, 1948, 14, 117-124.

21. Hemelrijk, J., A family of parameterfree tests for symmetry with respect to a given point I and II. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A), 1950, 53, 945-955, 1186-1198.

22. Hemelrijk, J., Note on Wilcoxon's two-sample test when ties are present. Annals of Mathematical Statistics, 1952, 23, 133-135.

23. HODGES, J. L. and LEHMANN, E. L., The efficiency of some nonparametric competitors of the t-test. Annals of Mathematical Statistics, 1956, 27, 324-335.

24. Keeping, E. S., The problem of birth ranks. Biometrics, 1952, 8, 112-119.

25. Kemperman, J. H. B., De verdelingsfunctie van het aantal inversies in de test van Mann en Whitney, Mimeographed report No. T. W. 7, Mathematical Centre, Amsterdam, 1950.

26. Kempthorne, O., The design and analysis of experiments, New York: Wiley, 1952, pp. 128-132.

27. Kempthorne, O., The randomization theory of experimental inferrence. Journal of the American Statistical Association, 1955, 50, 946-967.

28. Kendall, M. G., The advanced theory of statistics, London: Griffin, 1946, Vol. II, pp. 122-127.

29. Kruskall, W. H., Historical notes on the Wilcoxon unpaired two-sample tests. Journal of the American Statistical Association, 1957, 52, 356-360.

30. Kruskall, W. H. and Wallis, W. A., Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association. 1952, 47, 583-621.

31. Kruskall, W. H. and Wallis, W. A., Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 1953, 48, 907-911.

32. Lehmann, E. L., Consistency and unbiasedness of certain nonparametric tests. Annals of Mathematical Statistics, 1951, 22, 165-179.

33. Lehmann, E. L., The power of rank tests. Annals of Mathematical Statistics, 1953, 24, 23-43.

34. Lehmann, E. L. and Stein, C., On the theory of some nonparametric hypotheses. *Annals of Mathematical Statistics*, 1949, 20, 28-45.

T* 35. MANN, H. B. and WHITNEY, D. R., On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947, 18, 50-60.

T 36. Mathematical Centre, *Auxiliary table for Wilcoxon's two sample test,* Report No. R132/S86, Mathematical Centre , Amsterdam.

37. Mood, A. M., On the asymptotic efficiency of certain nonparametric two-sample tests. *Annals of Mathematical Statistics*, 1954, 25, 514-522.

38. Nair, K. R., The median in tests by randomization. *Sankhya*, 1940, 4, 543-550.

39. Neyman, J., Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, 1942, 105, 292-327.

40. Pearson, E. S., Some aspects of the problem of randomization. *Biometrika,* 1937, 29, 53-64.

41. PITMAN, E. J. G., Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society,* (Series B), 1937, 4, 119-130.

* 42. Pitman, E. J. G., Significance tests which may be applied to samples from any populations II. The correlation coefficient test. *Journal of the Royal Statistical Society (Series B)* 1937, 4, 225-232.

* 43. Pitman, E. J. G., Significance tests which may be applied to samples from any populations III. The analysis of variance test. *Biometrika*, 1937, 29, 322-335.

44. Putter, J., The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics*, 1955, 26, 368-386.

T* 45. van der Reyden, D., A simple statistical significance test. Rodesia Agricultural Journal, 1952, 49, 96-104.

T 46. Rijkoort, P. J., Nomogram betreffende de toets van Wilcoxon, Nomogram published by the Royal Netherlands Meteorological Institute, De Bilt, Holland.

47. Ruist, E., Comparison of tests for non-parametric hypotheses. Arkiv für Matematik, 1954, 3, 133-163.

48. Scheffé, H., Statistical inference in the non-parametric case. Annals of Mathematical Statistics, 1943, 14, 305-332.

49. Stoker, D. J., An upper bound for the deviation between the distribution of Wilcoxon's test statistic for the two-sample problem and its limiting normal distribution for finite samples I, II. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A), 1954, 57, 599-606, 607-614.

50. Stuart, A., The asymptotic relative efficiencies of tests and the derivatives of their power functions. Skandinavisk Aktuarietidskrift, 1954, , 163-169.

51. Sundrum, R. M., A further approximation to the distribution of Wilcoxon's statistic in the general case. Journal of the Royal Statistical Society (Series B), 1954, 16, 255-260.

52. Sundrum, R. M., The power of Wilcoxon's 2-sample test. Journal of the Royal Statistical Society (Series B), 1953, 15, 246-252.

T 53. TUKEY, J. W., The simplest signed-rank tests, Statistical Research Group, Princeton University, Memorandum Report - 17.

54. van der Vaart, H. R., A closed expression for certain probabilities in Wilcoxon's two sample test. Separatum Experientia, 1956, 12, 14.

55. van der Vaart, H. R., An investigation on the power function of Wilcoxon's two sample test if the underlying distributions are not normal. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* (Series A), 1953, 56, 438-448.

56. van der Vaart, H. R., *De toets van Wilcoxon*, (English summary) Published lecture notes for material published also in (59).

57. van der Vaart, H. R., On a basic distribution-free multidecision solution of a certain k-sample problem. (Abstract) *Proceedings of the International Mathematical Congress*, Amsterdam, Sept. 1954.

58. van der Vaart, H. R., On certain statistical methods used in biology with special reference to Husson's paper on *cricetus cricetus canescens* nehring. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen*, 1953, 56, 631-638.

59. van der Vaart, H, R., Some remarks on the power function of Wilcoxon's test for the problem of two samples. I and II. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A)*, 1950, 53, 494-506, 507-520.

60. van der Waerden, B. L., Order tests for the two-sample problem and their power. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A)*, 1952, 55, 453-458.

61. van der Waerden, B. L., Order tests for the two-sample problem and their power (corrigenda). *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A)*, 1953, 56, 80.

62. van der Waerden, B. L., Order tests for the two-sample problem II. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A)*, 1953, 56, 303-310.

63. van der Waerden, B. L., Order tests for the two-sample problem III. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (Series A)*, 1953, 56, 311-316.

T 64. Walsh, J. E., Applications of some significance tests for the median which are valid under very general conditions. Journal of the American Statistical Association, 1949, 44, 342-355.

T 65. Walsh, J. E., On a generalization of the Behrens-Fisher problem. Human Biology, 1950, 22, 125-135.

T* 66. WALSH, J. E., Some significance tests for the median which are valid under very general conditions. Annals of Mathematical Statistics, 1949, 20, 64-81.

67. Welsh, B. L., On tests for homogeneity. Biometrika, 1938, 30, 149-158.

68. Welsh, B. L., On the Z-test in randomized blocks and latin squares. Biometrika, 1937, 29, 21-52.

T* 69. WHITE, C., The use of ranks in a test of significance for comparing two treatments. Biometrics, 1952, 8, 33-41.

TT** 70. WILCOXON, F., Individual comparisons by ranking methods. Biometrics, 1945, 1, 80-83.

T* 71. Wilcoxon, F., Individual comparisons of grouped data by ranking methods. Journal of Economic Entomology, 1946, 39, 269.

TT 72. Wilcoxon, F., Probability tables for individual comparisons by ranking methods. Biometrics, 1947, 3, 119-122.

TT 73. WILCOXON, F., Some rapid approximate statistical procedures, American Cynamid Company pamphlet, 1949.

74. Wilcoxon, F., Some rapid approximate statistical procedures. Annals of the New York Academy of Science, 1950, 52, 808-814.

75. WILKS, S. S., Order Statistics. Bulletin of the American Mathematics Society, 1948, 54, 6-50.

See also textbooks and other general sources in Introduction, and see k-sample tests for extension of Wilcoxon test to analysis of treatment effects in more than two samples.

CHAPTER VI

TESTS BASED ON THE METHOD OF RANDOMIZATION II

Fisher's method can be applied to almost any type of statistic or sample information. In the present chapter it is extended to testing for correlation, the most significant such application being that in which the method is used to obtain exact tables for Spearman's rank difference correlation coefficient.

1. Pitman's Correlation Test

a. Rationale. Suppose that an x observation and a y observation have been made on each of n units or individuals and that Pearson's product moment correlation coefficient, r, has been calculated from the data in the usual way. Now suppose that the correlation coefficient is calculated for every possible set of paired xs and ys, using the same data but permitting any given x observation to be paired with any of the n y observations, not just the one recorded for the same unit. There are n ways of assigning a y to x_1, n-1 ways of assigning a y to x_2 after making the first assignment, etc., so that there are in all n! ways of re-pairing the xs and ys. Let N be the number of these ways which result in a correlation coefficient as large or larger than that obtained for the data as recorded. If there is no correlation between x and y in the sampled population, then each of the n! correlation coefficients is equally likely and the a priori probability of obtaining a correlation coefficient as great or greater than that actually obtained is N/n!

b. Null Hypothesis. Each of the n! sets of pairs of xs and ys is equally likely to have been recorded. This will be the case if all assumptions are true and if there is no correlation between x and y.

c. Assumptions. Sampling is random, pairs of observations are independent and the sampled populations are continuously distributed so that there are no tied observations.

d. Treatment of Ties. If any xs or ys are tied there will be less than n! distinguishable sets of pairs. However, if ties are due to imprecision of measurement, the tied observations may be treated as if distinguishable, by regarding one tied observation as "green", another as "yellow", a third as "red", etc., in permuting data, so that n! remains the proper denominator for the probability fraction. To minimize error, half of the sets of pairs which, because of ties, yield exactly the same r as the actually recorded data may be counted as among the N "as extreme or more extreme" sets. For a conservative test, N should include all of them.

e. Efficiency. No information available.

f. Application. Let the obtained data be represented as follows:

						Sum	Mean
x	1	2	5	8	14	30	6
y	2	1	7	10	15	35	7
xy	2	2	35	80	210	$\sum xy = 329$	

The expression for r is $\dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2\sum(y - \bar{y})^2}}$ whose numerator is

$\sum xy - \bar{x}\sum y - \bar{y}\sum x + \sum \bar{x}\bar{y}$ or $\sum xy - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\,\bar{x}\bar{y}$ or simply

$\sum xy - n\bar{x}\bar{y}$ and whose denominator remains constant for every set of re-paired xs and ys. The N "most extreme" rs therefore will be those which have the N "most extreme" numerators. The numerator for the observed data is 329 - 210 or 119. This value can be exceeded in only one way: by switching the two leftmost ys. Therefore for a one-tailed test of the null hypothesis that there is either zero or negative correlation, $\alpha = 2/5! = 2/120$. For a two-tailed test N must include those sets of pairs for which the numerator of r is -119 or less, i.e., those sets for which $\sum xy \leq 210 - 119 = 91$. In this particular case there are no such sets, so the significance level for a two-tailed test is still $\alpha = 2/120$.

g. Discussion. In common with other tests based on Fisher's Method of Randomization and using original continuously distributed, measurements, this test is a conditional one. Strictly speaking statistical inference can be extended only to a "population" consisting of the xs and ys actually recorded, not to the larger population from which they were drawn. To the extent that the obtained sample is representative or typical of the larger population, it would be legitimate to extend inference to the larger population. However, such representativeness is not tested by the test and remains an unproven assumption for which there is generally little or no evidence. Likewise, because the rejection region varies with the sample, it is impossible to construct generally useful tables of probabilities for the test.

h. Tables. There are no tables; probabilities must be calculated for each individual case.

i. Sources. 23, See also 1.

2. The Rank Difference Correlation Coefficient

a. Rationale. If an x measurement and a y measurement have been taken on each of n units or individuals, the Pearson product moment correlation coefficient is

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}.$$ However, if the measurements are continuously distributed so that there are no ties, and if each measurement is replaced by its rank among measurements of the same type,

the formula for Pearson's r reduces to $r = 1 - \frac{6\sum d^2}{n^3 - n}$ where d is the difference between ranks of measurements taken on the same unit (13). The latter formula is the expression for Spearman's rank difference correlation coefficient, ρ. Therefore if original measurements are replaced by their ranks, Pitman's test, applying Fisher's Method of Randomization to the product moment correlation coefficient, and the application of Fisher's Method of Randomization to Spearman's rank difference correlation coefficient are mathematically equivalent. By using ranks, however, instead of original measurements, the test is no longer conditional upon the particular measurements recorded, the sample space and the rejection region for the test statistic are the same from one test to another for the same sample size n, and significance levels may profitably be tabled.

Therefore, let each x measurement be replaced by its rank among the xs, and each y measurement by its rank among the ys.

There are n! ways of obtaining a sample of n pairs of ranks, each pair containing an x rank and a y rank. If there is no correlation between x and y, each of these n! samples was equally likely, on an a priori basis, to have been the obtained sample. Therefore, if N of these n! samples yield a rank difference correlation coefficient as extreme or more extreme than that calculated for the actually obtained sample, the probability for that of the obtained sample is N/n!

b. Null Hypothesis. Each of the n! sets of pairs of xs and ys is equally likely to have been recorded. This will be the case if all assumptions are true and if there is no correlation between x and y.

c. Assumptions. Sampling is random, pairs of observations are independent, and the sampled populations are either continuously distributed or are natural rank populations consisting of the unrepeated integers from 1 to n so that there are no tied ranks.

d. Treatment of Ties. When the same value is recorded for more than one x observation or for more than one y observation, the problem of ties is raised. It has generally been recommended that such ties be given the midrank for the tied group in which they appear. However, Thornton (31) has pointed out when n "is very small one or more pairs of tie rankings will change very greatly the frequencies with which various values of $\sum d^2$ and ρ can be obtained", and has questioned "whether tie rankings tend to increase the probability of positive coefficients and to decrease the probability of negative coefficients." A perfect positive correlation, +1, is obtained when $\sum d^2 = 0$. For an n of 3, this can occur in the following ways if ties are assigned the midrank.

x	1	2	3	1 1/2	1 1/2	3	2	2	2	1	2 1/2	2 1/2
y	1	2	3	1 1/2	1 1/2	3	2	2	2	1	2 1/2	2 1/2

A perfect negative correlation, -1, is obtained when $\sum d^2 = \frac{n^3 - n}{3}$.

For an n of 3, the required sum of squared differences, 8, can occur only for the case of no ties:

x	1	2	3
y	3	2	1
d	-2	0	2
d^2	4	0	4

If any two of the three xs or of the three ys are tied, the corresponding d^2s will sum to less than 4 and the total sum of d^2s will be less than 8.

Such considerations suggest that the most reasonable treatment of ties is to distribute the tied-for ranks among the tied observations in each group in that way which is least conducive to rejection of the null hypothesis. The limits of "tie error" can be obtained by calculating probabilities under both the above method and the method by which tied-for ranks are assigned to tied observations in the way most conducive to rejection.

e. Efficiency. Spearman's rank difference correlation coefficient has an asymptotic estimate efficiency of $9/\pi^2$ or .912 as an estimator of Fearson's product moment correlation coefficient when the latter is zero and when both coefficients are obtained from large samples from a bivariate normal population (13). Under the conditions outlined above, therefore, the rank difference test for correlation has an asymptotic relative efficiency of .912 relative to the parametric test for correlation (27, 26).

The test has been shown by Hoeffding (10, 11) to be asymptotically biassed for certain alternatives.

f. Application. Using the same data used in the example of application of Pitman's correlation test we have:

x	1	2	5	8	14	
y	2	1	7	10	15	
x rank	1	2	3	4	5	
y rank	2	1	3	4	5	
d	-1	+1	0	0	0	
d^2	1	1	0	0	0	$\sum d^2 = 2$

The value of ρ for the obtained sample is $\rho = 1 - \dfrac{6 \sum d^2}{n^3 - n} = 1 - \dfrac{6 \text{ x } 2}{125-5}$

$= .90$ which can be exceeded in only one way — by switching the y ranks 1 and 2 so as to obtain a perfect positive correlation. It can be equalled, however, by any one of the following four ways, the x ranks being listed only once since re-pairing can be accomplished by manipulating only the ys:

x	1	2	3	4	5
y	2	1	3	4	5
y	1	3	2	4	5
y	1	2	4	3	5
y	1	2	3	5	4

Therefore for a one-tailed test of the hypothesis that correlation is either zero or negative, $\alpha = N/n! = \dfrac{4+1}{5!} = \dfrac{5}{120}$. For a two-tailed test, N must include all sets of re-paired ranks which yield a ρ of -.90 or a larger negative magnitude. They are listed as follows:

$$\left.\begin{array}{cccccc} x & 1 & 2 & 3 & 4 & 5 \\ y & 5 & 4 & 3 & 2 & 1 \end{array}\right\} \qquad \begin{array}{l} \sum d^2 = 40 \\ \rho = -1 \end{array}$$

$$\left.\begin{array}{cccccc} x & 1 & 2 & 3 & 4 & 5 \\ y & 4 & 5 & 3 & 2 & 1 \\ y & 5 & 3 & 4 & 2 & 1 \\ y & 5 & 4 & 2 & 3 & 1 \\ y & 5 & 4 & 3 & 1 & 2 \end{array}\right\} \qquad \begin{array}{l} \sum d^2 = 38 \\ \rho = -.90 \end{array}$$

It is clear therefore that the test statistic is symmetrically distributed so that the significance level for a two-tailed test is just twice that for a one-tailed test, i.e., $\alpha = N/n! = 10/120$. In actual application, of course, the significance levels would be obtained directly from tables rather than by enumerating the number of ways which constitute the numerator of the probability fraction.

g. Discussion. It has been forcefully pointed out (22 and editorial note accompanying 30) in the past that correlation between sets of ranked variate values is not the same thing as correlation between sets of original variate measurements. Recent results by Stuart, however, indicate that when samples are of moderate or large size, conclusions as to correlation among original measurements may reasonably be drawn from tests of correlation which use only the ranks. Stuart (27, see also 15 pp. 124-125) found that when sample size increased from 25 to infinity the correlation between original measurements and their ranks increased from .94 to .98, for samples from normally distributed populations, and from .96 to 1.00 for samples from uniformly distributed populations with finite range.

There are, at present, two outstanding rank tests for correlation, the present test and Kendall's rank order test of correlation. The two tests are not mathematically equivalent: "it is possible to have populations in which $\tau = 0$ and $\rho = 1/2$ or $-1/2$" (4). However, when applied to samples from bivariate normal populations in which x and y are uncorrelated, Spearman's ρ and Kendall's τ are highly correlated. For such cases, the product moment correlation coefficient for correlation between ρ and τ is .980 when n = 5, .990 when n = 20, and 1.00 when n = ∞ (15 p. 80, 6, 5).

When applied to very large samples from a bivariate normal population in which the population product moment correlation between x and y is $\hat{r}$, the product moment correlation between Spearman's ρ and Kendall's τ is 1 when $\hat{r}$ = 0, .9996 when $\hat{r}$ = .2, .9981 when $\hat{r}$ = .4 and .9843 when $\hat{r}$ = .8, "though it tends to zero as $\hat{r}$ approaches unity" (15, p. 131).

The rank difference correlation test has the advantage that it can be performed very quickly. Also, because rank differences are squared, the test is particularly desirable when one wishes to weight large discrepancies between ranked xs and ys more heavily than small ones. In most other respects, however, the test appears to be inferior to Kendall's rank order correlation test (15, 16, 18).

Both the distribution of ρ and that of τ approach the normal distribution as n increases (13, 10, 5). However, the distribution of ρ is inadequately approximated by the normal distribution when samples are of a size just too large for the exact tables, which extend from n = 2 to n = 10, to be applicable. The "fit" between the distribution of ρ and its normal approximation is poor at the most important region, the tails, when n is small, e.g. when n = 11. Furthermore, at these small sample sizes the distribution of ρ is very jagged ordinatewise, presenting a sawtoothed appearance (15, 16). By contrast, the distribution of Kendall's τ approaches the normal form much more rapidly so that the normal approximation is reasonably good at those sample sizes at which it must be used to obtain probabilities. At these sample sizes the distribution of τ is such that the curve descends monotonically on either side of its mode, the entire curve including its tails giving the appearance of a very nearly normal distribution (15).

A modification of the rank difference correlation test has been considered by Daniels (4) as a test for trend. It has an asymptotic relative efficiency of $(3/\pi)^{1/3}$ or .98, relative to the regression coefficient test, b, as a test of randomness against normal regression alternatives. When applied in these circumstances, it is equal in efficiency to Mann's T test, and generally superior to other distribution-free tests of randomness (29, 26). See Table I of the Introduction.

h. Tables. Exact probabilities have been tabled for $2 \leqq n \leqq 7$ by Olds (20), for $2 \leqq n \leqq 8$ by Kendall, Kendall and Smith (16), for $n = 9$ and $n = 10$ by David, Kendall and Stuart (6), and for $4 \leqq n \leqq 10$ by Kendall (15). Approximate probabilities have been tabled for $8 \leqq n \leqq 30$ by Olds (20, 21) using a Type II curve for $8 \leqq n \leqq 10$ and using the normal approximation for $11 \leqq n \leqq 30$. All of these tables are entered with $\sum d^2$ rather than ρ. Thornton (31) has "translated" Olds' tables of probabilities for $\sum d^2$ into probabilities for ρ. Olds' tables have been criticized as containing distortions when sample sizes are in the region of $n = 11$ (31).

If there is no correlation between ranked xs and ys, then as n increases the sampling distribution of ρ approaches a normal distribution whose mean is zero and whose variance is $\frac{1}{n - 1}$. Likewise, the sampling distribution of $\sum d^2$ approaches a normal distribution whose mean is $\frac{n^3 - n}{6}$ and whose variance is $\left(\frac{n^3 - n}{6}\right)^2 \frac{1}{n - 1}$.

Therefore, for large samples, $\frac{\rho}{\sqrt{n - 1}}$ or $\dfrac{\sum d^2 - \frac{n^3 - n}{6}}{\frac{n^3 - n}{6\sqrt{n - 1}}}$ may be treated as normal deviates with zero mean and unit variance, and probabilities may be obtained by referring these critical ratios to normal tables. Various corrections to these formulae are available which "correct" for the effect of ties (12, 15, 30, 36) or for discontinuity (See 15, pp. 34-35, 38-41, 59-60). However, because of the biassing effect produced by ties, the most reasonable procedure would appear to be the most conservative one. Following this philosophy tied observations would be assigned the tied-for ranks least conducive to rejection of the null hypothesis. Probabilities would then be obtained using formulae "uncorrected" for ties. When there are no ties, the interval between successive values of $\sum d^2$ is 2, so the appropriate correction for continuity consists of subtracting or adding 1 to the numerator of the critical ratio. If the numerator is positive, it should be decreased by 1, if negative, increased by 1.

i. Sources. 2-18, 20-22, 25-31, 35, 36.

3. Test for Serial Correlation

Wald and Wolfowitz (32) have considered the Method of Randomization as a means of testing the significance of the serial correlation coefficient,

$$R_h = \frac{\sum_{i=1}^{n} X_i X_{i+h} - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

There are n! permututations of the order in which the Xs were actually recorded, and for all permutations $\sum_{i=1}^{n} X_i$ (and $\sum_{i=1}^{n} X_i^2$) will be the same. Therefore, the statistic used is simply $R_h = \sum_{i=1}^{n} X_i X_{i+h}$, the subscript i indicating the i^{th} X in order of appearance and h indicating the "lag" or indicating the period of a suspected cyclical fluctuation. When $i+h > n$, X_{i+h-n} is used instead of X_{i+h}. The value of R_h is calculated, in effect, for each of the n! possible permutations of order (which are equally probable if the null hypothesis, that the Xs are independent observations from the same population, and therefore appear in random order, is true). The N "values which constitute the critical reginn will depend in each particular problem on the possible alternatives to randomness," and so will the value of h. The significance level is N/n!, and the null hypothesis is rejected if the actually obtained value of R_h is among the N values of R_h which constitute the critical region. It is assumed that the Xs come from a continuously distributed population.

The value R_1 is asymptotically normally distributed (under mild qualifications) and, if h is prime to n, the distribution of $R_h = \sum_{i=1}^{n} x_i x_{i+h}$ is the same as the distribution of $R_1 = \sum_{i=1}^{n} x_i x_{i+1}$. Therefore, by taking h and n so that h is prime to n, the significance of R_h can be tested, for large samples, by referring the critical ratio

$$\frac{R_1 - \bar{R}_1}{\sigma_{R_1}}$$

to normal tables. Unfortunately, considerable calculation is required to obtain the mean and variance of R_1.

The authors have suggested that the test might be improved if the Xs were replaced by their ranks. Noether (19) finds that both tests, i.e., the one already outlined and the one in which Xs are replaced by their ranks, are consistent against certain alternatives of cyclical trend where h is the length of cycle. He finds that either test may have the greater asymptotic relative efficiency with respect to the other, depending upon the distribution of the population of Xs. The asymptotic relative efficiency of the R_h test relative to Mann's T test was found by Noether to be zero under certain stated conditions. It also has A.R.E. of zero relative to the best parametric test based on the regression coefficient (29, 26).

BIBLIOGRAPHY

1. Baten, W. D. and Trout, G. M., A critical study of the summation-of-difference-in-rank method of determining proficiency in judging dairy products. Biometrics, 1946, 2, 67-69.

2. Cureton, E. E., Rank-biserial correlation. Psychometrika, 1956, 21, 287-290.

3. Daniels, H. E., Note on Durbin and Stuart's formula for E(r_s). Journal of the Royal Statistical Society, (B), 1951, 13, 310.

* 4. Daniels, H. E., Rank correlation and population models. Journal of the Royal Statistical Society (B), 1950, 12, 171-181.

5. Daniels, H. E., The relation between measures of correlation in the universe of sample permutations. Biometrika, 1943, 33, 129-135.

T 6. David, S. T., Kendall, M. G. and Stuart, A., Some questions of distribution in the theory of rank correlation. Biometrika, 1951, 38, 131-140.

7. Durbin, J. and Stuart A., Inversion and rank correlation coefficients. Journal of the Royal Statistical Society (B), 1951, 13, 303-309.

8. Eells, W. C., Formulas for probable errors of coefficients of correlation. Journal of the American Statistical Association, 1929, 24, 170-173.

9. Ehrenberg, A. S. C., On sampling from a population of rankers. Biometrika, 1952, 39, 82-87.

10. Hoeffding, W., A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics, 1948, 19, 293-325.

11. Hoeffding, W., A non-parametric test of independence. *Annals of Mathematical Statistics*, 1948, 19, 546-557.

12. Horn, D., A correction for the effect of tied ranks on the value of the rank difference correlation coefficient. *Journal of Educational Psychology*, 1942, 33, 686-690.

* 13. HOTELLING, H. and PABST, MARGARET R., Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, 1936, 7, 29-43.

14. Kendall, M. G., Rank and product-moment correlation. *Biometrika*, 1949, 36, 177-193.

T 15. KENDALL, M. G., *Rank correlation methods*, 2nd Ed., New York: Hafner, 1955.

16. Kendall, M. G., Kendall, Shelia F. H. and Smith, B. B., The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, 1938, 30, 251-273.

17. Lyerly, S. B., The average Spearman rank correlation coefficient. *Psychometrika*, 1952, 17, 421-428.

18. Moran, P. A. P., Rank correlation and product-moment correlation. *Biometrika*, 1948, 35, 203-206.

19. Noether, G. E., Asymptotic properties of the Wald-Wolfowitz test of randomness. *Annals of Mathematical Statistics*, 1950, 21, 231-246.

T 20. Olds, E. G., Distributions of sums of squares of rank differences for small numbers of individuals. *Annals of Mathematical Statistics*, 1938, 9, 133-148.

T 21. Olds, E. G., The 5% significance levels for sums of squares of rank differences and a correction. *Annals of Mathematical Statistics*, 1949, 20, 117-118.

22. Pearson, K. and Heron, D., On theories of association, *Biometrika*, 1913, 9, 159-315.

* 23. Pitman, E. J. G., Significance tests which may be applied to samples from any populations II. The correlation coefficient test. Journal of the Royal Statistical Society, (B) 1937, 4. 225-232.

* 24. Pitman, E. J. G., Significance tests which may be applied to samples from any populations III. The analysis of variance test. Biometrika, 1937, 29, 322-335.

25. Spearman, C., The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101.

26. Stuart, A., Asymptotic relative efficiences of distribution-free tests of randomness against normal alternatives. Journal of the American Statistical Association, 1954, 49, 147-157.

27. STUART, A., The correlation between variate-values and ranks in samples from a continuous distribution. British Journal of Statistical Psychology, 1954, 7, 37-44.

28. Stuart, A., The correlation between variate-values and ranks in samples from distributions having no variance. British Journal of Statistical Psychology, 1955, 8, 25-27.

29. Stuart, A., The efficiencies of tests of randomness against normal regression. Journal of the American Statistical Association, 1956, 51, 285-287.

30. Student, An experimental determination of the probable error of Dr. Spearman's correlation coefficients. Biometrika, 1921, 12, 263-282.

T 31. Thornton, G. R., The significance of rank difference coefficients of correlation. Psychometrika, 1943, 8, 211-222.

* 32. Wald, A. and Wolfowitz, J., An exact test for randomness in the non-parametric case based on serial correlation. Annals of Mathematical Statistics, 1943, 14, 378-388.

33. Welsh, B. L., On tests for homogeneity. *Biometrika,* 1938, 30, 149-158.

34. Welsh, B. L., On the Z-test in randomized blocks and Latin squares. *Biometrika*, 1937, 29, 21-52.

35. Whitefield, J. W., Uses of the ranking method in psychology. *Journal of the Royal Statistical Society, (B)* 1950, 12, 163-170.

36. Woodbury, M. A., Rank correlation when there are equal variates. *Annals of Mathematical Statistics*, 1940, 11, 358-362.

CHAPTER VII

TESTS BASED UPON INVERSIONS

Correlation can be tested by arranging units in increasing order of one variable and testing the resulting order of the other variable for randomness. If there are n units and there is no correlation, the resulting sequence of observations on the second variable is equally likely to be any of the n! possible permutations of the observations. However, if the two variables are linearly correlated, the observations on the second variable should tend to form an increasing or decreasing sequence, and the number of inversions in this sequence should tend to be extreme. By using the number of inversions as test statistic and applying essentially Fisher's Method of Randomization, an exact test for correlation can be formed and its probabilities tabled. By taking "time" as the first variable, the test can be made a test for trend.

1. The Distribution of Inversions

Let the integers from 1 to n be arranged in some order, such as the following: 3 5 1 4 2 6. When a given number is followed by a smaller number an inversion exists. In the sequence of integers just presented, there are six inversions: 3 is followed by two smaller numbers 1 and 2; 5 is followed by three smaller numbers 1, 4 and 2; and 4 is followed by the smaller number 2.

If the order in which the n integers are to be arranged is determined by a random process, then each of the n! permutations of the n integers is equally probable. And the a priori probability of obtaining a random sequence with exactly I inversions is simply the number of permutations containing exactly I inversions divided by n!, the number of permutations possible.

Besides I, two additional measures directly related to inversions will be encountered. For a single permutation the maximum number of inversions is simply the number of pairs of integers which are compared, $\binom{n}{2}$ or $\frac{n}{2}(n-1)$. Therefore the number of times an integer is followed by a larger integer in the sequence is the compliment of I and is equal to $\frac{n}{2}(n-1) - I$. This measure will be designated as T. The other measure is S which is equal to T - I. The following table gives the distributions of I, T and S for n = 4.

The distribution of I has mean $\frac{n}{4}(n-1)$ and variance $\frac{n(n-1)(2n+5)}{72}$, and as n approaches infinity it approaches the normal distribution (8, 40, 54). Therefore, for large n the critical ratio

$$\frac{I - \frac{n}{4}(n-1)}{\sqrt{\frac{n(n-1)(2n+5)}{72}}}$$ may be treated as a normal deviate.

TABLE V

DISTRIBUTION OF INVERSIONS FOR n = 4

Permutation	I	T	S	Permutation	I	T	S
1 2 3 4	0	6	6	3 1 2 4	2	4	2
1 2 4 3	1	5	4	3 1 4 2	3	3	0
1 3 2 4	1	5	4	3 2 1 4	3	3	0
1 3 4 2	2	4	2	3 2 4 1	4	2	-2
1 4 2 3	2	4	2	3 4 1 2	4	2	-2
1 4 3 2	3	3	0	3 4 2 1	5	1	-4
2 1 3 4	1	5	4	4 1 2 3	3	3	0
2 1 4 3	2	4	2	4 1 3 2	4	2	-2
2 3 1 4	2	4	2	4 2 1 3	4	2	-2
2 3 4 1	3	3	0	4 2 3 1	5	1	-4
2 4 1 3	3	3	0	4 3 1 2	5	1	-4
2 4 3 1	4	2	-2	4 3 2 1	6	0	-6

I	T	S	Frequency	Probability	Cumulative Probability
0	6	6	1	1/24	1/24
1	5	4	3	3/24	4/24
2	4	2	5	5/24	9/24
3	3	0	6	6/24	15/24
4	2	-2	5	5/24	20/24
5	1	-4	3	3/24	23/24
6	0	-6	1	1/24	24/24
			24		

2. Kendall's Rank Order Correlation Test

a. Rationale. Suppose that an x measurement and a y measurement have been taken on each of n units and that tied xs and tied ys are both impossible. If the units are arranged from left to right in order of increasing x scores, the sequence of ys will be random if x and y are uncorrelated. However, if x and y are linearly correlated, the sequence of ys will tend to increase or decrease systematically, and the number of inversions among the ys will tend to be small or large respectively, Therefore the number of inversions among the ys can be used to test the null hypothesis that x and y are randomly associated against the alternative that they are linearly correlated.

Let the xs be ranked from 1 to n and the ys also, and let the units be arranged in increasing order of x rank. Then if T is the number of times a y rank is followed by a larger y rank and I is the number of times a y rank is followed by a smaller y rank, Kendall's test statistic is $S = T - I$. Since $T = \frac{n}{2}(n-1) - I$, $S = \frac{n}{2}(n-1) - 2I$ or $= 2T - \frac{n}{2}(n-1)$, so S, I and T are mathematically equivalent test statistics (when there are no tied scores).

The xs need not actually be arranged in order of increasing magnitude in order to calculate S. It is obvious from the foregoing that S is simply the number of the $\binom{n}{2}$ pairs of units in which the x and y scores of one member deviate in the same direction from their respective x and y counterparts in the other member minus the number of pairs in which they deviate in opposite directions. Therefore, let the units be arranged in any arbitrary order and let subscripts indicate position in this order, unit j being any unit to the right of unit i. Let a_{ij} be a dummy score which is +1 if x_j is greater than x_i and -1 if x_j is less than x_i. Similarly b_{ij} is +1 if $y_j > y_i$ and -1 if $y_j < y_i$. Finally, let $c_{ij} = a_{ij} b_{ij}$ so c_{ij} is +1 if $(x_i - x_j)(y_i - y_j)$ is positive, i.e., if either $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$, and is -1 if the product is negative. Then S is the sum of the c_{ij}s taken over all values of $j>i$ and all values of i from 1 to n.

b. Null Hypothesis. Each of the n! possible permutations of rank order of the ys was equally likely, before sampling, to be found when the units are arranged in order of increasing rank on the x measurement. A sufficient condition for the validity of the null hypothesis is that x and y are uncorrelated and all assumptions are true.

c. Assumptions. The units have been drawn independently and at random from a population in which each variable, x and y, is either continuously distributed or exists naturally in the form of untied ranks.

d. Treatment of Ties. If ties are due to imprecision of measurement, the safest rule is probably to distribute the tied-for ranks to the tied measurements in the way least conducive to rejection of the null hypothesis. The limits of tie-error can be obtained by comparing the probability obtained in this manner with that obtained by taking the opposite course. This rule may be safely followed regardless of whether exact or normal tables are used and without recourse to extensive corrections in formulae or to modifications of procedure. An alternative method is to give observations in each group of tied values the average of the ranks the members of the group would have if distinguishable. The midrank method, however, requires considerable qualification as will be shown in the paragraphs to follow.

When ties are assigned the midrank and probabilities are obtained from tables constructed upon the assumption that ties are impossible, the obtained probabilities are distorted; however, the statistic S is a far safer one than T or I. Consider first the case where ties are due to imprecision of measurement. If the ys are arranged in order of increasing x-rank and the last three of four y-ranks are tied, the y-ranks are 1 3 3 3, so $T = 3$, $I = 0$ and $S = 3$. The true ranking of the ys could be any of the following permutations:

Permutation	T	I	S
1 2 3 4	6	0	6
1 2 4 3	5	1	4
1 3 2 4	5	1	4
1 3 4 2	4	2	2
1 4 2 3	4	2	2
1 4 3 2	3	3	0

The average T is 4 1/2 but the value of T obtained by the midrank method lies at one end of the range of possible true values. The situation for I is analogous. However, the average S is precisely the value obtained by using the midrank. (The average S, however, is an odd number, whereas for n = 4 when there are no ties S assumes only even values. The probability tables therefore will have no entry for S = 3 and another source of inexactitude will have arisen.)

Consider now the case where ties represent intrinsic equality rather than imprecision of measurement. In this event, the proper tables are those based upon the frequency distribution of S given that certain ties exist, such as the tables prepared by Sillitto (43). The appropriate tables, therefore, would be derived by obtaining the frequency distributions of S when each ranking contains specified numbers of ties of specified extents, obtaining each such distribution by letting the y rankings assume every distinguishable permutation while the x ranking is held constant. The conventional tables, derived from untied rankings, are not appropriate and, if used in lieu of, or in the absence of, the proper tables, may lead to gross errors in probabilities. The amount of error attendant upon this procedure, however, is not the same for T, I and S. These three statistics are mathematically equivalent when ties are impossible, but not otherwise. When ties exist the maximum value T and I can assume is reduced, but the minimum value is the same as if ties were impossible. Since S is the difference between T and I, and since T is inversely related to I, S can assume neither the same maximum nor the same minimum as it could if ties were absent. The result is that the distributions of S when there are ties tend to maintain symmetry about the same point as that about which the distribution of S is symmetrical when ties are impossible. And since it is the extreme "tabled" values which become impossible when there are ties, the true probability of the central Ss tends to gain at the expense of the extremes. Therefore the error of referring S to tables based on the assumption of no ties is likely to be a decrease in the probability of rejection, and the error will tend to be a "conservative" one. Furthermore the error tends to be no greater for a one-tailed than for a two-tailed test. The distributions of T and I when there are ties tend to occupy a region closer to their minimal values than is the case when there are no ties. This distribution may be quite skewed, and even if it is symmetrical, the point of symmetry is closer to the minimal value than is the case when there are no ties. The result is that the true probability of the smallest values of T or I tend to be much greater than that obtained from tables based on no ties, thus spuriously increasing the probability of rejection when the rejection region consists of the smallest values. The situation is improved by using a two-tailed test, but the error may still be great in the direction of spurious rejection. The obvious conclusion is that, while there is no choice between T, I and S when ties are impossible, S is a much safer test statistic, although by no means free from error, when ties are present either because of imprecision of measurement or because of intrinsic equality.

If ties result from intrinsic equality between scores, the tie is not an artifact of measurement, but represents a fundamental discrepancy between the mathematical model and the situation it is intended to simulate. For such cases it is reasonable to alter the mathematical model. Sillito (43) has followed essentially this procedure by obtaining the exact distribution of S when there are ρ_2 groups of two tied scores and ρ_3 groups of three tied scores in one of the two rankings, the other ranking being tieless. He has tabled the probability of S for all possible values of ρ_2 and ρ_3 (and for all combinations thereof), from zero to the maximum number, for $3 \leqq n \leqq 10$. These probabilities are conditional probabilities: they state the probability of S given that one ranking is tieless and the other contains ρ_2 groups of two tied ranks and ρ_3 groups of three tied ranks. When there are ties and they are assigned the midrank, the mean of S remains zero, but its variance is altered. The formula for the variance of S when there are ties in one or both rankings has been obtained

TABLE VI

Conditional Frequency Distributions of T, I and S When n = 4 and There Are no Ties, One Tied Pair in One Ranking, Two Tied Pairs in One Ranking, and One Tie of Three Ranks in One Ranking

Frequency Distributions of T, I and S if:

Value of Statistic			Ties are Impossible	Two Ranks are Tied			Two Sets of Two Ranks are Tied			Three Ranks are Tied		
T	I	S	T, I or S	T	I	S	T	I	S	T	I	S
0	6	-6	1	3			1			2		
		-5				3						
1	5	-4	3	6	3		1		1	2		
		-3				6						2
2	4	-2	5	9	6		2	1	1	2		
		-1				9						2
3	3	0	6	9	9		1	1	2	2	2	
		1				9						2
4	2	2	5	6	9		1	2	1		2	
		3				6						2
5	1	4	3	3	6			1	1		2	
		5				3						
6	0	6	1		3			1			2	

TABLE VII

Cumulative Probability Distributions for T, I and S When n = 4 and There Are no Ties, One Tied Pair in One Ranking, Two Tied Pairs in One Ranking, and One Tie of Three Ranks in One Ranking

				True Distribution if								
Value of Statistic			"Tabled" Distribution: Ties are Impossible	Two Ranks are Tied			Two Sets of Two Ranks are Tied			Three Ranks are Tied		
T	I	S	T, I or S	T	I	S	T	I	S	T	I	S
0	6	-6	.04	.08			.17			.25		
		-5				.08						
1	5	-4	.17	.25	.08		.33		.17	.50		
		-3				.25						.25
2	4	-2	.38	.50	.25		.67	.17	.33	.75		
		-1				.50						.50
3	3	0	.63	.75	.50		.83	.33	.67	1.00	.25	
		1				.75						.75
4	2	2	.83	.92	.75		1.00	.67	.83		.50	
		3				.92						1.00
5	1	4	.96	1.00	.92			.83	1.00		.75	
		5				1.00						
6	0	6	1.00		1.00			1.00			1.00	
3±3	3±3	±6	.08	.08	.08		.17	.17		.25	.25	
		±5				.17						
3±2	3±2	±4	.33	.33	.33		.33	.33	.33	.50	.50	
		±3				.50						.50
3±1	3±1	±2	.75	.75	.75		.83	.83	.67	.75	.75	
		±1				1.00						1.00
3	3	0	1.00	1.00	1.00		1.00	1.00	1.00	1.00	1.00	

TABLE VIII

Conditional Frequency Distributions of T, I and S When n = 4 and There Are no Ties in Either Ranking and When n = 4 and There Are Three Tied Ranks in One Ranking and Either Two Tied Ranks, Two Sets of Two Tied Ranks, or Three Tied Ranks in the Other Ranking

			Frequencv Distribution of T, I and S if									
				One Ranking Contains Three Tied Ranks The Other:								
Value of Statistic			Ties are Impossible	Two Tied Ranks			Two Sets of Two Tied Ranks			Three Tied Ranks		
T	I	S	T, I or S	T	I	S	T	I	S	T	I	S
0	6	-6	1	24			6			8		
		-5										
1	5	-4	3	18						6		
		-3				12						2
2	4	-2	5	18		12	6		6			
		-1				6						6
3	3	0	6	12	12	12				2	2	
		1				6						6
4	2	2	5		18	12		6	6			
		3				12						2
5	1	4	3		18						6	
		5										
6	0	6	1		24			6			8	

TABLE IX

Cumulative Probability Distributions for T, I and S When n = 4 and There Are No Ties in Either Ranking and when n = 4 and There Are Three Tied Ranks in One Ranking and Either Two Tied Ranks, Two Sets of Two Tied Ranks, or Three Tied Ranks in the Other Ranking

Value of Statistic			"Tabled" Distribution: Ties are Impossible	True Distribution if One Ranking Contains Three Tied Ranks and the Other Ranking Contains:								
				Two Tied Ranks			Two Sets of Two Tied Ranks			Three Tied Ranks		
T	I	S	T, I or S	T	I	S	T	I	S	T	I	S
0	6	-6	.04	.33			.50			.50		
		-5										
1	5	-4	.17	.58						.88		
		-3				.17						.13
2	4	-2	.38	.83		.33	1.00		.50			
		-1				.42						.50
3	3	0	.63	1.00	.17	.58				1.00	.13	
		1				.67						.88
4	2	2	.83		.42	.83		.50	1.00			
		3				1.00						1.00
5	1	4	.96		.67						.50	
		5										
6	0	6	1.00		1.00			1.00			1.00	
3±3	3±3	±6	.08	.33	.33		.50	.50		.50	.50	
		±5										
3±2	3±2	±4	.33	.58	.58					.88	.88	
		±3				.33						.25
3±1	3±1	±2	.75	.83	.83	.67	1.00	1.00	1.00			
		±1				.83						1.00
3	3	0	1.00	1.00	1.00	1.00				1.00	1.00	

(23, 27, 43, 56. See also 52 for variance of T). Therefore when samples are large and ties are given the midrank, the significance of S can be obtained by referring the critical ratio, based upon the "corrected" variance, to normal tables. Again, however, the probability obtained is conditional upon the existence, in the population, (either the population of original scores or the corresponding population of measurements) of ties in precisely the number and extent implied by the corrected variance formula, e.g., in the same proportionate number and extent as exists in the obtained samples. Two further disadvantages are that the corrected variance formula is a long one and, when a critical ratio based upon it is referred to normal tables, the correction for continuity depends upon which of several tie situations exists and may not be precisely determinable.

When one takes a ranking containing ties, resolves the ties in all possible ways and calculates S for each way, the average of these Ss is the same as the S obtained by the midrank method. However, if, following Muhsam (36), one takes an untied ranking or pair of rankings, introduces all possible ties, calculates S for each case, and obtains the average S it is extremely unlikely to be the same as the S for the untied rankings, and the distribution of S's is likely to be quite skewed. In the former case one is dealing with a single conditional distribution since the number, extent and location of the tied groups is specified and fixed. In this case, if the null hypothesis is true, each of the untied rankings which might have been the true ranking is equally likely to have been obtained as an untied ranking and therefore should be equally likely to be the true "parent" of the tied ranking. In the latter case, however, the situation is quite different. One is dealing with a multiplicity of conditional distributions, e.g. in a ranking of four objects, the distribution of S conditional upon the existence of one tie of two rankings, the remaining two rankings being untied, or the distribution of S conditional upon the existence of one tie of four rankings, etc. To calculate S for all distinguishable rankings under all possible tying situations and then take the average S by summing the individual Ss and dividing by their total number is implicitly to assume that each such component S is equally probable. This in turn introduces the assumption that each tying condition's relative probability is in proportion to the number of distinguishable rankings which can be obtained from it. It can easily be shown that this assumption is false. In the example just given, there are $\frac{4!}{2!}$ distinguishable permutations of four ranks of which a specified two are tied, and these two can be any one of three pairs. So the total number of distinguishable permutations of four ranks, two of which are tied is $3(\frac{4!}{2!})$ or 36. However, all four ranks can be tied in only one way, so the ratio of the number of distinguishable rankings when there is one tie of two ranks to the number when there is one tie of four ranks is 36, i.e., the ratio is a constant for the case under consideration. Now let ρ be the unknown probability that a rank is tied with the "truly" next higher rank and let $q = 1 - \rho$ be the probability that it is not. A single tie of two ranks can be obtained in the following ways and with the following probabilities: rank 1 is tied with rank 2 but rank 2 is not tied with rank 3 and rank 3 is not tied with rank 4 ($Pr = \rho q^2$), rank 1 is not tied with rank 2, rank 2 is tied with rank 3, but rank 3 is not tied with rank 4 ($Pr=\rho q^2$); rank 1 is not tied with rank 2, rank 2 is not tied with rank 3, but rank 3 is tied with rank 4 ($Pr = \rho q^2$). The probability of a single tie of two ranks is therefore $3\rho q^2$. All four ranks can be tied in one way: if rank 2 is tied with rank 1 and rank 3 with rank 2 and rank 4 with rank 3 ($Pr = \rho^3$). The ratio of the probability for a tie of two to that of a tie of four is $\frac{3q^2}{\rho^2}$, i.e., is a variable which depends upon the unknown probability, ρ, that a rank will be tied with the "truly" next higher rank. Thus the distribution of S when all possible ties have been introduced in the rankings lacks meaning because the various rankings, so obtained, are not all equally probable when the null hypothesis is true.

e. Efficiency. When applied to samples of infinite size from bivariate normal populations in which x and y are uncorrelated, Kendall's tau is perfectly correlated with Spearman's rho (5, 9, 23). Therefore the asymptotic estimate efficiency of $9/\pi^2$ or .912 for rho as an estimator of Pearson's product moment correlation coefficient, when both coefficients are obtained from large samples from a bivariate normal population, applies equally to tau (47). Under the conditions stated, therefore, Kendall's rank order test for correlation has an asymptotic relative efficiency of .912 relative to the parametric test for correlation (17, 32).

The test has been shown to be consistent under conditions stated by Mann (29) and Terpstra (52). Conditions for its unbiassedness have been given by Mann (29).

f. Application. Designating units by letters, let the following data represent the variate values of x and y on each unit:

	UNIT	A	B	C	D	E
Measures	x	177	41	39	150	99
	y	84	4	7	53	16

Replacing variate values by their ranks, the data become:

	UNIT	A	B	C	D	E
Measures	x-rank	5	2	1	4	3
	y-rank	5	1	2	4	3

One method of calculating S does not require putting one ranking in the "natural" order, 1, 2, 3, etc. If there are ties in both rankings and if ties are given the midrank this method should be used.

For each of the $\binom{n}{2}$ possible pairs of units, a +1 is scored if the x and y of one unit deviate in the same direction from their respective counterparts of the other unit, i.e., if they are both higher or both lower than their counterparts in the other unit, otherwise, if the deviations are in opposite directions, a - 1 is scored. The sum of the $\binom{n}{2}$ plus or minus 1s is S. (If ties are given the midrank, the pairs for which the x ranks, the y ranks, or both, are tied are given a zero score.) For example, for the comparison involving units C and E the x rank of unit E is greater than the x rank of unit C and the y rank of unit E is also greater than the y rank of unit C. Therefore a score of +1 is recorded for this comparison. When unit C is compared with unit B the x rank of B is the greater of the two x ranks while its y rank is the lesser of the two, so a -1 is recorded. Of the ten possible comparisons of pairs of units, nine result in a score of +1 and one in a score of -1. So the algebraic sum S = +8.

If none of the x ranks are tied, the units may be arranged in order of increasing x rank thus simplifying the calculation of S for now for any pair of units a score of +1 results if the y rank of the unit "higher" in the series is greater than the y rank of the lower unit, and a score of -1 results in the opposite case. Rearranging the units so that the x ranks form an increasing sequence, the data appear as follows:

	UNIT	C	B	E	D	A
Measures	x-rank	1	2	3	4	5
	y-rank	2	1	3	4	5

And the calculation of S is shown in the following table:

y rank	Number of larger y ranks following	Number of smaller y ranks following
2	3	1
1	3	0
3	2	0
4	1	0
Sum	9	1

$S = 9 - 1 = +8$

The five y ranks can be permuted in 5! or 120 ways of which the following permutations result in an S as great or greater than +8:

1	2	3	4	5	S = +10
2	1	3	4	5	S = + 8
1	3	2	4	5	S = + 8
1	2	4	3	5	S = + 8
1	2	3	5	4	S = + 8

Therefore for a one-tailed test of the null hypothesis that the x and y ranks have either zero or negative correlation, the significance level is $\propto = 5/120$. For a two-tailed test of the hypothesis of zero rank correlation , the permutations yielding an S of - 8 or less must also be considered. They are:

5	4	3	2	1	S = -10
5	4	3	1	2	S = - 8
5	4	2	3	1	S = - 8
5	3	4	2	1	S = - 8
4	5	3	2	1	S = - 8

So $\propto = 10/120$ for the two-tailed test.

In actual application, the significance levels would be obtained from tables. Furthermore it is clear that the value of S can be obtained directly from the variate values of x and y without first converting these values into ranks. Conversion into ranks has two advantages however. It makes the counting process simpler and therefore reduces the likelihood of computational error. And it serves as a reminder that it is rank correlation which is being tested, not correlation among variate values. (The test could be used for the latter purpose, but only as a conditional test, i.e., its conclusions would be restricted to the set of observations obtained in the sample and could not be extended to the sampled population.)

g. Discussion. Kendall's rank order correlation test is one of the most important distribution-free tests. It is equalled in efficiency and excelled in speed of computation by Hotelling and Pabst's rank difference correlation test based on Spearman's rho. However, in most other respects it is the better test (23, 28, 34). (For a comparison of the two tests see Chapter VI.) As in the case of the test based on rho, a coefficient of correlation can be calculated from the data used in Kendall's test. The maximum value S can attain is simply the number of comparisons of pairs, $\binom{n}{2}$ or $\frac{n}{2}(n-1)$. Kendall therefore defines $\frac{S}{\frac{n}{2}(n-1)}$ to be his coefficient of rank correlation, which he calls tau. Its value ranges from -1 for perfect negative correlation to +1 for perfect positive correlation. Tau is related to rho, not directly, but by certain mathematical inequalities, e.g., $-1 \leq 3\tau - 2\rho \leq 1$. "When the sample is permuted in all possible ways", Daniels (5) finds the correlation between τ and ρ to be $\frac{2(n+1)}{\sqrt{2n(2n+5)}}$.

Kendall's statistic has many interesting properties. Moran (33) has shown that S is directly related to the "least number of interchanges of neighbors required to restore the permutation to the normal order", i.e., the order 1, 2, 3, ... , n-1, n. If i is the least number of such interchanges required, then $i = \frac{n(n-1) - 2S}{4}$.

A number of statistical tests may be regarded as the form which would be assumed by Kendall's test if it were modified to take account of ties representing intrinsic equality. For example, the Mann-Whitney test statistic U is the number of times an A-sample observation precedes a B-sample observation when the observations from both samples are arranged in order of increasing magnitude irrespective of sample. U therefore may be regarded as T with increasing rank order of magnitude for the combined sample as the x ranking and with the y ranking consisting of a set of tied" A's intermixed with a set of "tied" B's. In this case, the "ties" would be due to intrinsic equality, i.e., the "tied ranks" would define a category (sample A or sample B) and "rank" would have no quantitative meaning for the ys.

The conditional probability of T given that one ranking contained two sets of ties, one of extent a, the other, b = n - a, would therefore be identical to the probability of a U of the same value, U = T, when sample A contained a observations and sample B contained b observations, no observations being tied. It should be emphasized, however, that this conditional probability of T is obtainable from the U tables but not from the tables for T or S which are derived under the assumption of no tied ranks. That is to say, the proper tables for S, when there are ties, are those calculated from the conditional distribution of S given that so many groups of so many tied observations are present in the data. The situation is analogous when probabilities are obtained from the normal approximation. In that case the standard deviation used as the denominator of the critical ratio must be the square root of the conditional variance of S given that certain ties have occurred. The formulae for the "corrected" variance of S may become quite formidable as for instance in the case that there are ties in both rankings. Therefore, the relationship between Kendall's S "when there are ties" and other tests is a somewhat contrived one which is interesting but not particularly useful in most cases. Generally it will be more efficient and less confusing to employ tests expressly designed for data classified into groups or categories rather than to seek out the proper modification of the inversions test. This is especially true when samples are small since the exact conditional distribution of S apparently has been tabled only for the case where there are "ties" in one ranking and only then for $n \leqq 10$ (43).

Several tests have been developed which do not belong to the category discussed above. Whitfield (55) has outlined a test for intraclass correlation of ranked data and tabled its exact probabilities. Ranks from 1 to n are assigned to the members of n/2 pairs of observations. The pairs are then arranged in order of the lowest rank in each pair, i. e, the lowest rank among the remaining observations not yet ordered. Kendall's S is then calculated in the usual manner except that no observation is compared with its paired member (but is compared with all n-2 other observations). S max is therefore $\frac{n}{2}(n-2)$ and, since S min is zero, the average S is $\frac{n}{4}(n-2)$. Defining his test statistic as $S_\rho = S - \frac{n(n-2)}{4}$, Whitfield tables its probabilities for $6 \leqq n \leqq 20$. He finds its variance to be $\frac{n^3 - 4n}{18}$ so that large sample probabilities can be obtained by referring the critical ratio to normal tables. Moran (31) has outlined a curvilinear ranking test in which the integer 1 is moved to the nearest end of the range of ranks, then the integer 2 to the nearest end of the range of ranks 2 to n, etc., until all integers have been so treated. The test statistic is the leasi number of interchanges of integers required to effect this. Exact tables have been prepared for $2 \leqq n \leqq 14$. Daniels and Kendall (6) have developed a large sample test for the significance of the difference between two correlations when the correlations in the parent populations are not zero. They have also attacked the problem of establishing confidence limits for a rank correlation when a nonzero correlation exists in the parent population. Kendall (21) has established a partial correlation coefficient based on ranks, but has been unable to test its significance.

h. <u>Tables</u>. Probabilities for S have been tabled for $n \leqq 10$ by Kendall (23, 24) and for $n \leqq 40$ by Kaarsemaker and van Wijngaarden (19). (Because of the linear relation between S and T, the probability of the T corresponding to S is also the probability of S; therefore, T tables can also be used to test for rank correlation when there are no ties. See 3, Inversions as a Test for Linear Trend.)

If there are no ties and if all rank permutations are equally probable, i.e., if x and y are uncorrelated, the distribution of S rapidly approaches the normal distribution as n increases (5, 20, 23, 35, 44). Asymptotic normality of S in the null case has also been found when ties are present in one ranking (52) and, under certain conditions, when both rankings contain ties (7). When x and y are correlated, the distribution of S is asymptotically normal under certain stated conditions (16, 23).

When there are no ties, the distribution of S has mean zero and variance $n(n-1)(2n+5)/18$. Therefore when n is too large for the exact probability tables to be applicable, approximate probabilities can be obtained by referring the critical ratio $\dfrac{S}{\sqrt{\dfrac{n(n-1)(2n+5)}{18}}}$ to normal tables. The approximation can be improved by correcting for continuity. S is discretely distributed, successive values of S being two units apart; therefore, a tail area of the S distribution whose least extreme value is S would be represented, on a continuous curve, by an S one unit less extreme. The correction for continuity therefore consists of decreasing the value of S by one unit if S is positive or increasing it by one unit if it is negative, before calculating the critical ratio. If ties are present, different continuity corrections are required depending upon the situation. Some of these corrections have been given by Kendall (23).

The conditional variance of S given that certain ties exist and are assigned the midrank has been given by Kendall and others (23, 27, 43). If only one ranking contains ties, the variance of S is

$$\frac{n(n-1)(2n+5) - \sum t(t-1)(2t+5)}{18}$$ where t is the number of ranks in a tied group, i.e., the number of observations tied for a given value, the summation being taken over all such groups (the value of t perhaps varying from group to group). If both rankings contain ties, the variance of S is

$$\frac{n(n-1)(2n+5) - \sum t(t-1)(2t+5) - \sum \mu(\mu-1)(2\mu+5)}{18} +$$

$$\frac{\left\{\sum t(t-1)(t-2)\right\}\left\{\sum \mu(\mu-1)(\mu-2)\right\}}{9n(n-1)(n-2)} + \frac{\left\{\sum t(t-1)\right\}\left\{\sum \mu(\mu-1)\right\}}{2n(n-1)}$$ where t is defined as above but refers only to the ties in one ranking and μ is analogous to t but refers to ties in the other ranking. The mean S remains zero when there are ties in one or both rankings. When critical ratios are referred to normal tables, the proper correction for continuity depends upon the tying situation. The correction for certain cases has been given by Kendall (23).

j. Sources. 1-7, 9-11, 14-28, 31-36, 39, 42-44, 46-48, 50, 55-57.

3. Inversions as a Test for Linear Trend

a. Rationale. Suppose that, in Kendall's test for correlation, the x variable were the time at which a unit "appeared", or was generated, and the y variable were some quantitative measure on the unit itself. Kendall's test would then test whether or not the size of the y measurement is randomly related to the order in which the units were generated, and it would be particularly likely to reject the hypothesis of randomness if there were a linear trend in the generating process.

The test can be applied by following Kendall's procedure, in which case the test statistic is S and Kendall's tables are the appropriate ones to use, or by following a slightly different, but equivalent, procedure outlined by Mann. The observations, i.e. y measurements, are arranged in temporal order of appearance, and the number of times a subsequent measurement exceeds a given y is counted for each y and the sum obtained for all ys. This sum is called T. It is simply the complement of the number of inversions and is related to I and to S in the following manner $T = \frac{n}{2}(n-1) - I = S + I$.

b. Null Hypothesis. Each of the n! possible permutations of order for the size-rank of the ys was equally likely, before sampling, to result by arranging the ys in the temporal order in which they were generated. A sufficient condition for the validity of the null hypothesis is that the size of the y observations is uncorrelated with the temporal order in which they are generated and all assumptions are true.

c. Assumptions. The observations have been taken independently and at random from a population in which the ys are continuously distributed, or exist naturally in the form of untied ranks, and in which ys are generated one at a time.

d. Treatment of Ties. See 2, Kendall's Rank Order Correlation Test. If ties are given the midrank, S, rather than T or I, should be used as the test statistic.

e. Efficiency. When used as tests of randomness against normal regression alternatives, Mann's T test has asymptotic relative efficiency of $(3/\pi)^{1/3}$ or .98 relative to the parametric test based upon the regression coefficient, b (49, See also 45). It is therefore equal or superior to most other distribution-free tests for trend. See Table I in Introduction and (45, 49, 13, 37, 38).

The test is consistent and unbiassed (29, 14) under general conditions stated by Mann (29).

f. Application. If x is time in the example given under Application for Kendall's test, the time-ordered y values are 7, 4, 16, 53, 84 for which T = 9 and S = 8. Significance levels may therefore be obtained either by using Mann's probability tables for T or Kendall's for S.

g. Discussion. See 2, Kendall's Rank Order Correlation Test.

Elfving and Whitlock (12) have proposed a test for trend which is equivalent to T pooled over several sets of observations. The test statistic is equivalent to the sum of r Ts, where r is the number of sets of observations. Its mean and variance are the respective sums of the means and variances for the individual sets. Thus, in effect, the test is carried out by referring to normal tables a critical ratio whose numerator is $\sum_r [T - \frac{n(n-1)}{4}]$ and whose denominator is the square root of $\sum_r \frac{2n^3 + 3n^2 - 5n}{72}$, n referring to the number of observations in a set.

h. Tables. Mann (29) has tabled the exact T probabilities for $3 \leqq n \leqq 10$. By using S instead of T, Kendall's tables (23, 24), or the exact tables of Kaarsemaker and van Wijngaarden (19) can be used, the latter yielding exact probabilities for n's up to 40.

The distribution of T has mean $\frac{n}{4}(n-1)$ and variance $\frac{2n^3 + 3n^2 - 5n}{72}$ and approaches a normal distribution as n approaches infinity (29, 40, 54). Therefore when n is too large for the exact tables to apply, approximate probabilities for T may be obtained by referring the critical ratio

$$\frac{T - \frac{n}{4}(n-1)}{\sqrt{\frac{2n^3 + 3n^2 - 5n}{72}}}$$

to normal tables. (To correct for continuity, positive numerators should be decreased, and negative numerators increased, by 1/2). However, if ties exist and are given the midrank, neither the T tables nor the normal approximation to T should be used. Instead, the test should be carried out using Kendall's S as the test statistic.

j. Sources. 4, 8, 12, 13, 14, 19, 23, 24, 29, 30, 37, 38, 40, 41, 43, 45, 49, 52, 53, 54.

4. Mann's K-Test

a. Rationale. Mann (29) finds that "If P ($X_i > X_j$) increases rapidly with j-i, then another test is more powerful than the T-test." This test, the K-test, consists of arranging the observations in their order of appearance, X_o, X_1, X_2, . . . , X_{n-1} and finding "the smallest value of K for which the following set of inequalities is fulfilled:

$$X_o > X_K, X_o > X_{K+1}, \dots, X_o > X_{n-1}$$
$$X_1 > X_{K+1}, \dots, X_1 > X_{n-1}$$
$$\dots\dots\dots\dots\dots\dots$$
$$X_{n-K-1} > X_{n-1}"$$

The probability that for n untied observations K will be some specified integer $\bar{K}$ is simply the number of permutations in which K = $\bar{K}$ divided by n!, the number of possible permutations of the n observations.

b. Null Hypothesis. See 3, Inversions as a Test for Linear Trend.

c. Assumptions. See 3.

d. Treatment of Ties. Make no compromise in interpreting the inequality sign (see above) when determining K. The probabilities thus obtained will err in the conservative direction, i.e., rejection will be less likely than if there were no ties.

e. Efficiency. Mann states that when Pr ($X_i > X_j$) increases rapidly with j-i, the K-test is more powerful than the T-test. He notes "that the K-test is most powerful with respect to a fairly wide class of alternatives".

f. Application. In the example given for the T-test, the time-ordered observations are 7, 4, 16, 53, 84. There is obviously no downward trend, which the K-test is designed to detect. However, the presence of an upward trend may be tested by reversing the signs of the inequalities given under Rationale, and proceeding in the manner outlined for downward trend. The 7 is exceeded by all observations from the 16 on, the 4 is exceeded by each of the two observations following the 16, and the 16 is exceeded by the 84. Therefore K is the subscript which goes with 16, which is the third observation in order, therefore having the subscript 2, since subscripts start with zero. So K = 2 which for n = 5 has a tabled probability of .0667.

g. Discussion. This test has two outstanding disadvantages. First, it is easy to make errors in determining K. The determination of K involves examining several possibilities in order to pick the smallest K satisfying a rather involved set of inequalities. And the subscript notation is a confusing one since K is one unit less than the positional rank of the observation to which it refers. Furthermore, for certain order permutations there is no value of K which satisfies the inequalities. (Zero cannot be used to designate K in this situation because zero refers to the first observation in order of appearance.) Second, the K-test apparently is restricted to one-tailed tests of hypotheses. K is not symmetrically distributed, so the two-tailed probability cannot be obtained by doubling the one-tailed probability. And the value of K, for a test of downward trend, though different from the value it takes in testing for upward trend, presumably is not entirely independent of it. So, if the presumption is correct, two-tail probabilities cannot be obtained by combining probabilities from two opposite one-tailed tests. The following table serves to illustrate these points.

h. Tables. Mann (29) has tabled the probability that $K \leq \overline{K}$ for $3 \leq n \leq 9$. Actually these tables will suffice for almost any practical application regardless of the value of n. For n = 10 and K = 5, the first five observations are compared with the last five (i.e., the following comparisons are made: X_0 with X_5, X_6, X_7, X_8, X_9; X_1 with X_6, X_7, X_8, X_9; X_2 with X_7, X_8, X_9; X_3 with X_8, X_9; and X_4 with X_9). Since under the null hypothesis the observations are randomly arranged in order, for n > 10 the test may be arbitrarily applied to only the ten observations consisting of the first five and the last five in the series. When n = 10, the probability that $K \leq 5$ is .0098. In this case, this is also the probability that $K \leq n - 5$. When n is greater than 10 and $4 < K < n - 5$, if the set of inequalities holds for K, it will also hold for a "K" of n - 5 when the set of observations is reduced in size to incude only the first five and last five observations. The probability of the latter will be greater than that of the former, but the increase will be from some value smaller than .0098 to .0098, thus still being beyond the .01 level of significance. Therefore, for practical purposes only the first five and last fiv e observations are necessary to conduct a reasonable test of significance.

i. Sources. 29

TABLE X

TABULATION OF THE DISTRIBUTION OF K WHEN n = 4

Size-Rank Permutation				K = 1	K = 2	K = 3	K = Nothing
1	2	3	4				x
1	2	4	3				x
1	3	2	4				x
1	3	4	2				x
1	4	2	3				x
1	4	3	2				x
2	1	3	4				x
2	1	4	3				x
2	3	1	4				x
2	3	4	1			x	
2	4	1	3				x
2	4	3	1			x	
3	1	2	4				x
3	1	4	2			x	
3	2	1	4				x
3	2	4	1			x	
3	4	1	2		x		
3	4	2	1		x		
4	1	2	3			x	
4	1	3	2			x	
4	2	1	3			x	
4	2	3	1		x		
4	3	1	2		x		
4	3	2	1	x			
Point Probability				.042	.167	.292	.500
Cumulative Probability				.042	.209	.500	1.000

BIBLIOGRAPHY

1. Adler, Leta M., A modification of Kendall's tau for the case of arbitrary ties in both rankings. *Journal of the American Statistical Association*, 1957, 52, 33-35.

2. Cartwright, D. S., A note concerning Kendall's tau. *Psychological Bulletin*, 1957, 54, 423-425.

3. Cureton, E. E., Rank-biserial correlation. *Psychometrika*, 1956, 21, 287-290.

4. Daniels, H. E., Rank correlation and population models. *Journal of the Royal Statistical Society (B)*, 1950, 12, 171-181.

5. Daniels, H. E., The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 1943, 33, 129-135.

6. Daniels, H. E. and Kendall, M. G., The significance of rank correlations where parental correlation exists. *Biometrika*, 1947, 34, 197-208.

7. van Dantzig, D. and Hemelrijk, J., Statistical methods based on few assumptions. *Bulletin of the International Statistical Institute*, 1954, 34,

8. Dantzig, G. B., On a class of distributions that approach the normal distribution function. *Annals of Mathematical Statistics*, 1939, 10, 247-253.

9. David, S. T., Kendall, M. G. and Stuart, A., Some questions of distribution in the theory of rank correlation. *Biometrika*, 1951, 38, 131-140.

10. Durbin, J. and Stuart, A., Inversions and rank correlation coefficients. *Journal of the Royal Statistical Society (B)*, 1951, 13, 303-309.

11. Ehrenberg, A. S. C., On sampling from a population of rankers. *Biometrika*, 1952, 39, 82-87.

12. Elfving, G. and Whitlock, J. H., A simple trend test with application to erythrocyte size data. *Biometrics*, 1950, 6, 282-288.

13. Foster, F. G. and Stuart, A., Distribution-free tests in time-series based on the breaking of records. *Journal of the Royal Statistical Society (B),* 1954, 16, 1-13.

14. Hoeffding, W., A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics,* 1948, 19, 293-325.

15. Hoeffding, W., A non-parametric test of independence. *Annals of Mathematical Statistics*, 1948, 19, 546-557.

16. Hoeffding, W., On the distribution of the rank correlation coefficient τ when the variates are not independent. *Biometrika,* 1947, 34, 183-196.

17. Hotelling, H. and Pabst, Margaret R., Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, 1936, 7, 29-43.

18. Jones, M. B., An addition to Schaeffer and Levitt's "Kendall's tau". *Psychological Bulletin,* 1957, 54, 159-160.

T 19. KAARSEMAKER, L. and VAN WIJNGAARDEN, A., *Tables for use in rank correlation,* Report No. R 73 of the Computation Department, Mathematical Centre, Amsterdam.

* 20. KENDALL, M. G., A new measure of rank correlation. *Biometrika*, 1938, 30, 81-93.

21. Kendall, M. G., Partial rank correlation. *Biometrika,* 1942, 32, 277-283.

22. Kendall, M. G., Rank and product-moment correlation. *Biometrika*, 1949, 36, 177-193.

TT 23. KENDALL, M. G., *Rank correlation methods*, 2nd Ed., New York: Hafner, 1955.

TT 24. KENDALL, M. G., *The advanced theory of statistics, Vol. I,* London: Griffin, 1947, pp. 388-408.

25. Kendall, M. G., *The advanced theory of statistics, Vol. II,* London: Griffin, 1946, pp. 122-124.

26. Kendall, M. G., The treatment of ties in ranking problems. *Biometrika,* 1945, 33, 239-251.

27. Kendall, M. G., The variance of τ when both rankings contain ties. *Biometrika,* 1947, 34, 297-298.

28. Kendall, M. G., Kendall, Shelia, F. H. and Smith, B. B., The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika,* 1938, 30, 251-273.

*TT 29. MANN, H. B., Nonparametric tests against trend. *Econometrica,* 1945, 13, 245-259.

30. Mann, H. B., Nonparametric tests against trend. In *Statistical Inference in Dynamic Economic Models,* Ed. by T. C. Koopmans, New York: Wiley, 1950, pp. 345-351.

*T 31. Moran, P. A. P., A curvilinear ranking test. *Journal of the Royal Statistical Society (B),* 1950, 12, 292-295.

32. Moran, P. A. P., Rank correlation and permutation distributions. *Proceedings of the Cambridge Philosophical Society,* 1948, 44, 142-144.

33. Moran, P. A. P., Rank correlation and product-moment correlation. *Biometrika,* 1948, 35, 203-206.

34. Moran, P. A. P., Recent developments in ranking theory. *Journal of the Royal Statistical Society, (B),* 1950, 12, 153-162.

35. Moran, P. A. P., Partial and multiple rank correlation. *Biometrika,* 1951, 38, 26-32.

36. Muhsam, H. V., A probability approach to ties in rank correlation. *Bulletin of the Research Council of Israel,* 1954, 3, 321-327.

37. Noether, G. E., Asymptotic properties of the Wald-Wolfowitz test of randomness. *Annals of Mathematical Statistics*, 1950, 21, 231-246.

38. Noether, G. E., *Sequential tests of randomness*, Report No. OSR - TN - 54 - 65, under Contract AF 18(600)-778, Boston University.

39. Putter, J., The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics*, 1955, 26, 368-386.

40. Rosander, A. C., The use of inversions as a test of random order. *Journal of the American Statistical Association*, 1942, 37, 352-358.

41. Savage, I. R., Contributions to the theory of rank order statistics - the "trend" case. *Annals of Mathematical Statistics*, 1957, 28, 968-977.

42. SCHAEFFER, M. S. and LEVITT, E. E., Concerning Kendall's tau, a nonparametric correlation coefficient. *Psychological Bulletin*, 1956, 53, 338-346.

T 43. SILLITTO, G. P., The distribution of Kendall's τ coefficient of rank correlation in rankings containing ties. *Biometrika*, 1947, 34, 36-40.

44. Silverstone, H., A note on the cumulants of Kendall's S-distribution. *Biometrika*, 1950, 37, 231-235.

45. Stuart, A., Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *Journal of the American Statistical Association*, 1954, 49, 147-157.

46. Stuart, A., Bounds for the variance of Kendall's rank correlation statistic. *Biometrika*, 1956, 43, 474-477.

47. STUART, A., The correlation between variate-values and ranks in samples from a continuous distribution. *British Journal of Statistical Psychology*, 1954, 7, 37-44.

48. Stuart, A., The correlation between variate-values and ranks in samples from distributions having no variance. British Journal of Statistical Psychology, 1955, 8, 25-27.

49. Stuart, A., The efficiencies of tests of randomness against normal regression. Journal of the American Statistical Association, 1956, 51, 285-287.

50. Sundrum, R. M., Moments of the rank correlation coefficient τ in the general case. Biometrika, 1953, 40, 409-420.

51. Terpstra, T. J., A generalization of Kendall's rank correlation statistic II. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1956, 59, 59-66.

52. Terpstra, T. J., The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1952, 55, 327-333.

53. Terpstra, T. J., The exact probability distribution of the T statistic for testing against trend and its normal approximation. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1953, 56, 433-437.

54. Ville, J., Sur l'application, à un critere d'indépendance, du dénombrement des inversions présentées par une permutation. Comptes Rendues (Paris), 1943, 217, 41-42.

*T 55. Whitfield, J. W., Intra-class rank correlation. Biometrika, 1949, 36, 463-467.

* 56. Whitfield, J. W., Rank correlation between two variables, one of which is ranked, the other dichotomous. Biometrika, 1947, 34, 292-296.

57 Whitfield, J. W., Uses of the ranking method in psychology. Journal of the Royal Statistical Society (B), 1950, 12, 163-170.

CHAPTER VIII

RUNS OF CONSTANT PROBABILITY EVENTS

In a series of two kinds of events, a and b, although the proportionate number of a's and b's will necessarily depend upon the ratio of their individual constant probabilities of occurrence, the pattern in which the obtained a's and b's arrange themselves will not and will be random unless a's and b's are sequentially dependent. In that case like events may tend to cluster, and this may be indicated by an unusually small number of runs, or clusters of like objects, in the pattern, or by runs of unexpected length. Thus the total number of runs, the length of the longest run, and various other run statistics can be used as the sample information with which to test for randomness of pattern of arrangement against the alternative of sequential dependency. By judicious definition of the two types of event, this test can be employed to test whether two sampled populations are identical, whether a trend exists in a sequentially sampled population, whether learning is taking place, etc. Run tests are often rather weak and inefficient, depending upon the type of application contemplated. However, their power may be greatly increased by introducing certain modifications (such as Ramachandran and Ranganathan's) or by combining the run test with an independent test (as in David's Chi-square Smooth test of goodness of fit).

1. Basic Formulae

A run is an unbroken sequence of similar events or like objects. For example in the series a a b a b b b a a there are five runs: one run of a's of length 1, two runs of a's of length 2, one run of b's of length 1 and one run of b's of length 3. The following notation will be used in the derivation of run formulae when there are two kinds of objects. Let r_{ij} be the number of runs of objects of type i whose length is j and let r_i be the number of runs of objects of type i irrespective of length, i.e. of all lengths. Let n_i be the number of objects of type i and let n be the number of objects of both types. The two types of objects will be designated 1 and 2 respectively. The only things which can interrupt or terminate a run of like objects are a run of the other type objects or else termination of the entire series. Therefore r_1, the number of runs of 1's can either be one greater than, equal to, or one less than r_2, the number of runs of 2's. When $r_1 = r_2 + 1$ the series can begin (and end) in only one way - with a run of 1's. Likewise when $r_2 = r_1 + 1$ the series must begin and end with runs of 2's. However, when $r_1 = r_2$ the series can either begin with a run of 1's and end with a run of 2's, or begin with a run of 2's and end with a run of 1's. Therefore it will be convenient to introduce the notation $F(r_1, r_2) = 1$ if $r_1 \neq r_2$

$$= 2 \text{ if } r_1 = r_2.$$

The r_1 runs of 1's of various lengths can be permuted in $r_1!$ ways. But a permutation which merely exchanges the positions of runs of 1's of the same length does not change the appearance of the series. The r_{ij} runs of 1's of length j can be permuted in $r_{1j}!$ ways without changing the appearance of the series. Therefore, the number of distinguishable permutations of the r_1 runs of 1's is

$\dfrac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!}$. For each of these distinguishable permutations of runs of 1's, there are $\dfrac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!}$ distinguishable permutations of the runs of 2's. And since, if $r_1 = r_2$ the series can begin in two ways, the total number of distinguishable permutations given that there are r_1 runs of 1's and r_2 runs of 2's of specified lengths is

$$\frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \qquad \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!} \qquad F(r_1, r_2).$$

Finally, since there are $\frac{n!}{n_1!\, n_2!}$ dintinguishable permutations of n_1 1's and n_2 2's, the probability that there will be exactly r_{11} runs of 1's of length 1, r_{12} of length 2, etc., as well as exactly r_{21} runs of 2's of length 1, r_{22} of length 2, etc., given that there are n_1 1's and n_2 2's in the series is

$$\Pr(r_{ij}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \quad \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!} \quad \frac{F(r_1, r_2)}{n!/n_1!\,n_2!}$$

Suppose that we are interested in the breakdown of runs of 1's according to length, but that we are not interested in the corresponding breakdown of runs of 2's. Considering only the 1's, there are $\frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!}$ distinguishable permutations of the r_1 runs of 1's. Now imagine the n_2 2's arranged in a line. There are n_2-1 spaces between 2's, and the r_2 runs of 2's can be obtained by selecting r_2-1 of these n_2-1 spaces and "widening" them for occupation by runs of 1's. This can be done in $\binom{n_2-1}{r_2-1}$ ways. If $r_1 = r_2-1$, then any given permutation of runs of 1's can be fitted into a specified r_2-1 spaces between 2's in only one way, since the series must start and end with a run of 2's. If $r_1=r_2+1$, in addition to the r_2-1 spaces between 2's the runs of 1's also occupy the space to the left of the leftmost 2 and to the right of the rightmost 2., The series starts and ends with runs of 1's, and the r_2-1 spaces between 2-runs are occupied by the second to the r_1-1st 1-run. Again, this can be accomplished in only one way. However, if $r_1=r_2$, the first 1-run can be placed either to the left of the leftmost 2-run, or between the first and second 2-runs. Therefore the probability of exactly r_{11} runs of 1's of length 1, r_{12} of length 2, etc., and r_2 runs of 2's of any lengths given that there are n_1 1's and n_2 2's is

$$\Pr(r_{1j}, r_2) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \binom{n_2-1}{r_2-1} F(r_1, r_2) \Big/ \frac{n!}{n_1!\, n_2!}$$

Suppose now that we are interested in neither the lengths nor the total number of runs of 2's. The r_1 runs of 1's can be inserted into any r_1 of the n_2+1 spaces before, between, and after the 2's, i.e., into any of the n_2-1 spaces between 2's as well as the space to the left of the leftmost 2 and the space to the right of the rightmost 2. This can be done in $\binom{n_2+1}{r_1}$ ways. Therefore, (the rest of the derivation being analogous to that given earlier) the probability of exactly r_{11} runs of 1's of length 1, r_{12} of length 2, etc., given that there are n_1 1's and n_2 2's is

$$\Pr(r_{1j}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \binom{n_2+1}{r_1} \Big/ \frac{n!}{n_1!\, n_2!}$$

Since the number of runs of 2's is unspecified, it may be r_1-1, r_1 or r_1+1 and the term $F(r_1, r_2)$ is not required in the formula.

The preceding formulae give probabilities for the entire run pattern in the sense that the exact number of runs of each possible length is specified, at least for runs of one type. In order to obtain the more general probability for only certain specified r_{ij}, one fixes these r_{ij} as constants and sums the formula over all other values for which the relationships,

$$\sum_{j=1}^{n_i} j\, r_{ij} = n_i \text{ and } \sum_j r_{ij} = r_{i.}, \text{ are satisfied.}$$ For example if

$n_1 = 7$ and $n_2 = 9$ the probability of exactly one run of 1s of length 4 would be

$$\Pr(r_{14} = 1) = \sum \frac{r_1!}{r_{11}!\, r_{12}!\, r_{13}!\, 1!} \binom{10}{r_1} \Big/ \frac{16!}{7!\, 9!} =$$

$$\frac{\frac{4!}{3!\,1!}\binom{10}{4} + \frac{3!}{1!\,1!\,1!}\binom{10}{3} + \frac{2!}{1!\,1!}\binom{10}{2}}{\frac{16!}{7!\,9!}},$$

since a run of length 4 could be accompanied by three runs of length 1, one of length 1 and one of length 2, or by one of length 3 while still fulfilling the condition that $n_1 = 7$ and since the number of runs of 1s in these three cases is 4, 3, and 2 respectively.

Now suppose that we are interested in number of runs, only, and not in their lengths. Imagine the n_1 1s arranged in a line. There are $n_1 - 1$ spaces between 1s and the 1s can be separated into r_1 runs by selecting and "widening" $r_1 - 1$ of these spaces, then filling them with runs of 2s. The $r_1 - 1$ spaces can be selected in $\binom{n_1 - 1}{r_1 - 1}$ ways. For each of these ways the r_2 runs of 2s (which will eventually be interlaced with the 1s) can, by analogous reasoning, be selected in $\binom{n_2 - 1}{r_2 - 1}$ ways. Any given set of r_1 runs of 1s and r_2 runs of 2s can be fitted together in one way if $r_1 = r_2 \pm 1$ and in two ways if $r_1 = r_2$. The number of distinguishable permutations of r_1 runs of 1s and r_2 runs of 2s is therefore $\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}F(r_1, r_2)$. The number of distinguishable permutations of n_1 1s and n_2 2s without restriction as to numbers of runs is $\frac{n!}{n_1!\, n_2!}$. Therefore the probability of exactly r_1 runs of 1s and r_2 runs of 2s given that there are n_1 1s and n_2 2s is

$$\Pr(r_1, r_2) = \binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1} F(r_1, r_2) \Big/ \frac{n!}{n_1!\, n_2!}$$

If we are interested only in the number of runs of 1s and are indifferent to whether r_2 equals r_1-1, r_1 or r_1+1, we still select the r_1 runs of 1s by selecting r_1-1 of the n_1-1 spaces between 1s for widening. However, now the spaces before and after the 2s as well as the spaces between 2s are available for occupation by 1s because the number of runs of 2s is not fixed. Therefore there are n_2+1 spaces available for occupation by the r_1 runs, and they can be chosen in $\binom{n_2+1}{r_1}$ ways. The rest of the derivation is analogous to that described earlier. Therefore, the probability that there will be exactly r_1 runs of 1s given that there are n_1 1s and n_2 2s is

$$\Pr(r_1) = \binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1} \Big/ \frac{n!}{n_1!\, n_2!} .$$

All of the run formulae heretofore listed take n_1 and n_2 as given. They give probabilities conditional upon the existence of exactly n_1 1s and n_2 2s in the obtained sample. If one is interested in the arrangement of 1s and 2s but not in the probability of obtaining a 1 or a 2, the foregoing formulae are generally the appropriate ones. However if "1" and "2" are mutually exclusive outcomes of a binomial event, with probabilities ρ and q respectively of occurrence on a single trial, the experimenter may be interested in the compound probability that there will be n_1 1s and n_2 2s and that their arrangement will contain a specified configuration of runs. This compound probability is obtained by taking the product of the binomial probability, $\binom{n}{n_1} \rho^{n_1} q^{n_2}$, and whichever one of the probability formulae listed earlier gives the appropriate conditional probability for the specified configuration of runs.

The various formulae given above could be used as the bases for a variety of statistical tests of the hypothesis that 1s and 2s are arranged randomly. The particular formula used would depend upon the conditions taken as given and upon the alternative hypothesis against which one wished the test to be most sensitive. However, although a multiplicity of such tests are possible, calculations of probabilities generally become quite involved at any but the smallest sample sizes. Therefore, in the following sections only those tests will be described for which probabilities have been tabled.

2. The Wald-Wolfowitz Total Number of Runs Test

a. Rationale. Suppose that two samples have been drawn (randomly and independently), each from a continuously distributed population, and that one wishes to test whether or not the parent populations are identical. Let the sizes of the two samples be m and n and let their observations be designated as xs and ys respectively. Now arrange the m+n observations in increasing order of magnitude irrespective of the sample to which an observation originally belonged. Finally, label each such observation x or y depending upon the sample from which it came. If the two samples came from identical populations, then the pattern of arrangement of xs and ys is a random one since x and y are arbitrary labels attached to observations drawn randomly and independently from the "same" population. However, if the samples are from different populations, one would expect observations from the same sample to tend to cluster; so the total number of runs should tend to be less than the number expected on a purely chance basis.

Let U stand for the total number of runs of both xs and ys. Since the number of runs of xs can be one less than, equal to, or one greater than the number of runs of ys, U can be an even number in only one way, but can be an odd number in two ways. Substituting into the formula

$$\Pr(r_1, r_2) = \binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1} \; F(r_1, r_2) \Big/ \binom{n}{n_1} ,$$

if $r_1 = r_2 = r$, $\Pr(r_1, r_2) = 2\binom{n-1}{r-1}\binom{m-1}{r-1} \Big/ \binom{m+n}{m}$,

if $r_1 = r$ and $r_2 = r + 1$, $\Pr(r_1, r_2) = \binom{n-1}{r-1}\binom{m-1}{r} \Big/ \binom{m+n}{m}$,

and if $r_1 = r + 1$ and $r_2 = r$, $\Pr(r_1, r_2) = \binom{n-1}{r}\binom{m-1}{r-1} \Big/ \binom{m+n}{m}$.

Therefore, the probability that the total number of runs will be some even number, 2r, is $\Pr(U = 2r) = 2\binom{n-1}{r-1}\binom{m-1}{r-1} \Big/ \binom{m+n}{m}$ and the probability that it will be some odd number, 2r+1 is

$$\Pr(U = 2r+1) = \frac{\binom{n-1}{r-1}\binom{m-1}{r} + \binom{n-1}{r}\binom{m-1}{r-1}}{\binom{m+n}{m}} .$$

b. <u>Null Hypothesis.</u> Given that there are m xs and n ys, each of the $\binom{m+n}{m}$ distinguishable arrangements of xs and ys was equally likely to have been the arrangement actually obtained. A sufficient condition for the validity of the null hypothesis is that the x observations and y observations were drawn from identical populations and that all assumptions are true.

c. Assumptions. For each sample the observations were drawn randomly and independently from a continuously distributed population.

d. Treatment of Ties. Ties are a problem only when observations from both samples are tied for the same position, or rank, in order of increasing magnitude. In many, but not all, such cases the resolution of ties can affect the total number of runs. A tie for which the total number of runs varies depending on how the tie is broken will be a called a "critical" tie. For a conservative test critical ties should be resolved in the manner least conducive to rejection of the null hypothesis. However, if one wishes to minimize the average error in probabilities, the following method of dealing with critical ties may be pursued. For tied groups consisting of a single x and a single y, randomly select one-half of the groups and resolve ties so that the x precedes the y with which it is tied; for the remaining half, resolve ties so that the y precedes the x; if an odd group remains, resolve the ties by flipping a coin. For tied groups consisting of a single x and two ys, resolve ties so that for a randomly selected 1/3 of these groups the order is xyy, for another randomly selected 1/3 it is yxy, and for the remaining 1/3 it is yyx, any remaining groups being resolved by randomly selecting one of the orders xyy, xyx, yyx, a different randomly-selected order being used for each such group. To generalize: if there are k groups in which s xs and t ys are tied with one another, resolve ties by successively selecting $\binom{s+t}{s}$ of the k groups and replacing each of them with a different, randomly assigned one of the $\binom{s+t}{s}$ distinguishable orderings of s xs and t ys; if k is not divisible by $\binom{s+t}{s}$ resolve ties in the remaining groups by randomly assigning each of them a different one of the $\binom{s+t}{s}$ possible orderings.

e. Efficiency. When applied to symmetrical populations known to be equal in all respects except for location, a test for identical populations is equivalent to a test for equal means. When both tests are applied to samples from normally distributed populations with equal variances, the Wald-Wolfowitz form of the run test has relative to Student's t-test an asymptotic relative efficiency of zero (33 see also qualifications stated in 30, 33) and a small sample efficiency which, when each sample contains five or less observations, generally exceeds .96 and may be as high as .995 (13). It also has an A.R.E. of zero relative to the F ratio when applied to normal populations as a test for dispersion (33). The test compares poorly with other distribution-free tests (see Table I in Introduction). It had the least power of the tests investigated by van der Waerden (47), Epstein (14), and Lehmann (30), the former two authors sampling from normal populations with homogeneous variances, the latter sampling from a continuously distributed population. It was found by one or more of these authors to be inferior in power to the following tests: Student's t, van der Waerden's X-test, Lehmann's most powerful test, Mann-Whitney test, Westenberg's Median test, Epstein's exceedances test, Smirnov's maximum deviation test. The Wald-Wolfowitz test is consistent if the ratio m/n of sample sizes remains constant as sample sizes m and n approach infinity and if certain other very mild conditions are met (48, 29). If the ratio m/n does not remain constant, but approaches zero or infinity, the test is inconsistent. That is to say, if one sample is of much greater size than the other, observations from the sample of smaller size are almost certain to be separated from each other by observations from the larger sample; thus, the number of runs will tend to be a maximum regardless of whether the null hypothesis is true or false (29).

The power function for Steven's form of the run test has been obtained against the alternative of a Markov chain by David (10).

f. Application. Suppose that a sample of observations has been taken on randomly selected and assigned subjects under each of two treatments and that it is desired to test whether the two treatments differ in any measured respect. The data are shown below.

Treatment x	5	14	23	61	114	125	131
Treatment y	47	55	64	66	71		

If the data are arranged in order of increasing magnitude with the sample from which each observation came listed below it, we have:

5	14	23	47	55	61	64	66	71	114	125	131
x	x	x	y	y	x	y	y	y	x	x	x

There are three runs of xs and two of ys, so U = 5. Entering the probability tables for runs with m = 7, n = 5, and taking the smallest numbers of runs as the rejection region, we find that the largest value of U significant at the one-tailed .05 level of significance is 3. Since U = 5 in the above data, the hypothesis of identical populations, and therefore equal treatment effects, cannot be rejected at the significance level chosen. Since a casual inspection of the data strongly suggests that the populations have unequal variances, the above example serves to illustrate the weakness of the test.

g. Discussion. The total number of runs can be used as a test statistic in ways other than that described for the Wald-Wolfowitz form of the test. Actually the total number of runs is an appropriate test whenever one is interested in the randomness of arrangement of mutually exclusive events, fixed in number, and constituting a dichotomy. It can be used as a test for trend by labeling observations above and below the median as x and y respectively; if there is a linear trend, the number of runs should be smaller than that expected by chance. It can be used (19, 7) to test the randomness of wet and dry days in order of appearance; or to test whether occupied seats at a lunch counter tend to occur in isolation, bordered by vacant seats(15). In such cases the null hypothesis is simply that given m xs and n ys each of the $\binom{m+n}{m}$ distinguishable arrangements is equally probable. The assumptions are that there are only two mutually exclusive and unconfusable categories and that sampling is random and independent. The efficiencies found for the Wald-Wolfowitz test relative to Students t are, of course, not applicable here. The formulae for Pr(U), given under Rationale, apply in all of the above cases. One additional case in which it does not apply is that in which the m 1s and n 2s are arranged around a circle rather than in a straight line. Stevens (42) has derived the probability for the total number of runs in this case.

h. Tables. Probabilities for U have been tabled by Swed and Eisenhart (44) for $m \leqq n \leqq 20$, and for certain other cases. Major portions of their tables are republished in (I-8, I-23, I-43,); smaller portions can be found in (22, 23, 50). David (9) has provided tables appropriate when $m+n \leq 14$ and $2 \leq U \leq 14$.

The mean and variance of U are $\frac{2mn}{m+n} + 1$ and

$\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}$ respectively and U is asymptotically normally distributed if the ratio of sample sizes remains constant while sample sizes approach infinity (48). Therefore, when samples are too large for the tables to apply, approximate probabilities can be obtained by treating U as a normal deviate and referring the critical ratio

$$\frac{U - \frac{2mn}{m+n} - 1}{\sqrt{\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}}}$$

to normal tables. (To correct for continuity, reduce the absolute value of the numerator by 1/2.) Generally the test will be one-tailed with "too few" runs constituting the critical region, in which case, of course, a one-tailed probability must be read from the normal tables for the critical ratio.

i. Sources. 9, 10, 13, 14, 15, 22, 23, 29, 30, 33, 34, 35, 42, 44, 47, 48, 49, 50, 51, 52.

3. Length of the Longest Run

a. Rationale. Just as the total number of runs is an index of a possible tendency for like objects to cluster, so is the length of the longest run. Using the notation of Section 1, the probability that the longest run of 1s will be of length S can be obtained by taking n_1 and n_2 as fixed and summing the formula

$$\Pr(r_{1j}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1(S-1)}!\, r_{1S}!} \binom{n_2+1}{r_1} \Big/ \binom{n}{n_1}$$

over all values of r_1 and over all sets of $r_{11}, r_{12}, \ldots r_{1(S-1)}, r_{1S}$

which satisfy $\sum_{j=1}^{n_1} j\, r_{1j} = n_1$, $\sum_{j} r_{1j} = r_1$, and $r_{1S} \geqq 1$ and such that r_1 exceeds neither $n_1 - S + 1$ nor $n_2 + 1$. The probability that the longest run of either 1s or 2s will be of length S can be obtained by an analogous attack upon the formula

$$\Pr(r_{ij}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1S}!} \quad \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2S}!} \; F(r_1, r_2) \Big/ \binom{n}{n_1}$$

with the proviso that r_{1S} and r_{2S} cannot both be zero at the same time. The above method is involved and considerably more convenient formulae have been derived for such probabilities (1, 34, 38, 49); however, their derivation is not as simple as those which have been presented here.

b. Null Hypothesis. Given that a sample contains n_1 1s and n_2 2s, each of the $\binom{n_1+n_2}{n_1}$ distinguishable arrangements of 1s and 2s was equally likely to have been obtained prior to sampling.

c. Assumptions. Sampling is random, observations are independent, and all observations can be unmistakably classified into one of two mutually exclusive and unconfusable categories.

d. Treatment of Ties. Ties are a problem only when their resolution may change the length of the longest run. Such ties should be resolved in the manner least conducive to rejection of the null hypothesis or else dealt with in a manner analogous to that outlined for critical ties in Section 2, The Wald-Wolfowitz Total Number of Runs Test.

e. Efficiency. Power functions were obtained by Bateman (1) for the length of longest run and for the total number of runs as tests of randomness against the alternative of a simple Markov chain, i.e., that each event is dependent upon the preceding event but no other. For this case the length of longest run test was found to be less powerful than the total number of runs test.

f. Application. In the following series a a b b a a a a b b b b b b b a b a a a, the longest run contains 7 like objects. Referring to tables of probabilities with $n_1 = 10$, $n_2 = 10$ and longest run = 7, the probability that at least one run of length 7 or more will occur either among the a's or among the b's is found to be .032. The probability that at least one run of 7 or more b's will occur is .017.

g. Discussion. See 2, The Wald-Wolfowitz Total Number of Runs Test.

h. Tables. Bateman (1) has provided probability tables for "at least one greatest run, of either kind of element, of given length" for values of $n_1 + n_2 \leqq 20$. These are point probabilities, i.e., are for one or more runs exactly S in length. Mosteller (38) has tabled the probability of at least one run of length S or greater among elements of one type, either type or each type for $n_1 = n_2 = 5$, 10, 15, 20 or 25. Portions of Mosteller's tables have been republished by (50, I-15).

i. Sources. 1, 34, 38, 49, 50.

4. Length of Longest Run as a Test for Randomness against Trend Alternatives

Suppose that a series of observations have been taken upon a continuously distributed variable and that they have been arranged in the order in which they were drawn, no two observations having been drawn simultaneously. If each observation is now labeled A or B depending upon whether it is above or below the median for the entire series, the presence or absence of trend can be tested by using as the test statistic one of the following: the length of the longest run on one side, either side or both sides of the median. If there are an odd number of observations one of them will be the median and it should be discarded. This test has been proposed by Mosteller (38) who has published appropriate tables for the cases where $n_1 = n_2 = 5, 10, 15, 20$ or 25. See also (50, I-15).

5. Length of Longest Run in Binomial Trials

a. Rationale. Rationale of 3, Length of Longest Run, discussed the method of obtaining the formula for the probability that the longest run of 1s will be of exactly length S. This probability was obtained by taking n_1 and n_2 as fixed constants, and is contingent upon their having the values assigned them. Let Pr $(S | n_1)$ stand for such a probability, and let $n = n_1 + n_2$ be fixed. Now suppose that the occurrence of a 1 or a 2 is a binomial event with probability ρ or q respectively for a single trial. If, for every possible value of n_1, Pr $(S|n_1)$ is multiplied by $\binom{n}{n_1} \rho^{n_1} q^{n_2}$ and the products are summed, the sum is simply the a priori probability that in n binomial trials the longest run of consecutive 1s will be of exactly length S. More convenient methods and formulae are used in actual tabulation of probabilities (21, 34, 46, 49).

b. Null Hypothesis. The probability that in n trials there will be exactly n_1 1s is $\binom{n}{n_1} \rho^{n_1} q^{n_2}$ and for any obtained value of n_1 each of the $\binom{n_1+n_2}{n_1}$ distinguishable arrangements of 1s and 2s is equally probable.

c. Assumptions. Sampling is random; observations are independent; 1 and 2 are mutually exclusive outcomes of a binomial event with constant probabilities ρ and $q = 1 - \rho$ for a single trial.

d. Treatment of Ties. Break ties in the manner least conducive to rejection of the null hypothesis.

e. Efficiency. No information available.

f. Application. An experimenter wishes to test whether or not a monkey can learn to associate a red light with food. The monkey's food is always hidden in one of five boxes and the "reward" box is always illuminated by a red light. The probability of "success" on a single trial is therefore 1/5 if the null hypothesis of no learning is true. Consulting Grant's tables (20) the experimenter finds that when $\rho = 1/5$ a run of 4 or more successes in 40 trials is significant at the .05 level. Therefore he decides to run not more than 40 trials and to reject the null hypothesis whenever the number of consecutive successes reaches 4. The monkey's successes and failures to go first to the red-illuminated box are: F F F S F F F F F S S F S S S S, so only 16 of the maximum of 40 trials had to be run. The significance level, however, is not reduced but remains .05 since it had originally been intended to run as many as 40 trials if necessary.

g. Discussion. The question arises as to which type of test is appropriate, that which treats n_1 and n_2 as given or that which treats ρ as given. Mosteller's test for trend takes $n_1 = n_2 = \frac{n}{2}$ and indeed this must be the case since n continuously distributed observations are being classified as above or below their own median. In this case it would be very improper to treat "above the median" as a binomial event with probability 1/2 since in n trials of such an event, n_1 should be able to assume any value from zero to n, which is obviously impossible if n_1 is the number of the n observations above the median of the same n observations. Similarly if one were interested in the randomness of a seating arrangement, one would take the observed number of occupied and unoccupied seats as given since it is only the pattern of occupancy, not the probability of occupancy, in which one is interested.

On the other hand suppose that one knows that he is dealing with a binomial event (which is free to occur any number of times from zero to n in n trials) and that one can state, a priori, the exact value of the constant parameter ρ. Then by using the "binomial" approach outlined under Rationale one need only conduct that number of trials between S and some predetermined value, n, necessary to produce the criterion of S consecutive successes. Research efficiency has therefore been gained. Furthermore, when used as a test for learning, as outlined by Grant (20, 21) and as conducted under "Application", the "binomial" approach has particularly desirable features, i.e., the test is particularly sensitive to the alternative hypothesis. When learning begins ρ (which is constant only if the null hypothesis of no learning is true) increases. This causes n_1 to tend to assume a value greater than chance would have given it. And naturally with a greater number of successes there are more ways of obtaining a run of S consecutive successes and the probability of a run of length S increases simply because of the "inflated" value of n_1. Learning, however, also increases the probability that successful trials will be temporally adjacent. Therefore, learning makes rejection particularly likely by increasing both the probability of temporal association among the number of successes occurring and by tending to increase the number of successes beyond what would be expected if the null hypothesis were true.

h. Tables. Grant (20, 21) has tabled the probability of a run of at least S successes in n trials for the following values of ρ : 1/2, 1/3, 1/4, 1/5. See also (18).

i. Sources. 4, 5, 7, 8, 15,18, 19, 20, 21, 34, 39, 46, 49.

6. The Sum of Squared Run Lengths

The Wald-Wolfowitz total number of runs test is one of the least powerful distribution-free tests for goodness of fit, i.e, that two samples were drawn from identical populations. Presumably this is partly because the total number of runs does not directly take account of the lengths of runs which are the more explicit indices of the tendency of like objects to cluster. The length of the longest run, by taking account of only the longest run, ignores the "information" contained in the lengths of the less-than-longest runs. And in the case investigated by Bateman (1) this statistic was found to be less powerful than the total number of runs.

Ramachandran and Ranganathan (40) have proposed a test which overcomes the objections voiced above. Their test statistic, N, is the sum of the squares of lengths of runs, i.e., N = $\sum_j j^2 r_{1j} + \sum_j j^2 r_{2j}$. Thus all runs are taken account of, but each run is permitted to influence the test statistic in proportion to the square of its length. Its authors recommend the test for the same situation dealt with by Wald and Wolfowitz, i.e, observations are arranged in increasing order of magnitude and runs of Sample 1 observations and of Sample 2 observations are noted, the test being used to decide whether the two samples belong to identical, continuously distributed, populations. The authors, considering only the case where $n_1 = n_2$, have tabled values of N required for various levels of significance. The tabled values of N are exact for the cases $3 \leqq n_1 \leqq 5$ and approximate for $6 \leqq n_1 \leqq 15$, in the latter case having been obtained by reading points from a Type VI curve fitted to the true distribution of N.

7. Dixon's Test

A test analogous to that of Ramachandran and Ranganathan was proposed earlier by Dixon (12). Two samples of sizes m and n, with n < m, are drawn from continuously distributed populations and arranged in order of increasing magnitude irrespective of sample. There are n + 1 spaces between, before and after the n observations into which the m observations may be distributed. If the two samples are from the same population, one would expect the proportion of the m observations actually falling into a specified space to be $\frac{1}{n+1}$. Therefore Dixon subtracts the observed, proportion $\frac{m_i}{m}$, where m_i is the number of such observations actually falling in the i^th space, from the expected proportion $\frac{1}{n+1}$, and squares the difference. This is done for each value of i from 1 to n+1. The sum of these n+1 squared differences is taken as the test statistic and called c^2. Probability tables are provided for c^2 for cases in which neither m nor n is greater than 10. For larger values of m or n approximate probabilities can be obtained by a procedure which relates c^2 to the chi square distribution. For details see (12).

The quantity m_i is of course the length of the run of observations from the sample of size m which occupies the ith interval "between" observations from the other sample of size n. However, since the ith interval may be unoccupied, m_i may be zero. Therefore the quantity squared by Dixon, i.e., $\frac{1}{n+1} - \frac{m_i}{m}$ is not directly comparable to the quantity squared by Ramachandran and Ranganathan, i.e., the length of an actually obtained run which therefore cannot be zero. Another way of putting it is that while the value $\frac{1}{n+1}$ is the expected proportion of m-sample observations falling in the ith interval, it is not the average length of obtained, m-sample, runs.

Still another test somewhat similar to the two discussed above, as well as to the Mann Whitney test has been outlined by Mathen. See (32).

8. David's Chi Square "Smooth" Test of Goodness of Fit

One of the classic criticisms of the chi square test of goodness of fit is that, since deviations from expected values are squared before being divided by the expected value and summed, the test does not take account of the directions of deviations. For example, consider the following table in which the columns, from left to right, represent the corresponding, successive, abscissa intervals.

f_o	15	14	13	12	11	9	8	7	6	5
f_e	10	10	10	10	10	10	10	10	10	10
f_o-f_e	5	4	3	2	1	-1	-2	-3	-4	-5

If the only restraint is that $\sum f_o = \sum f_e$, then there are 9 degrees of freedom and the obtained value of 11 for X^2 has a probability of about .30. Although there is a strong indication that the left portion of the true curve lies above, and the right portion lies below, the hypothesized curve, chi square ignores this information and, dealing only with the magnitudes of the deviations, falls short of significance.

David (9) has proposed a test which takes account of both the magnitude and the direction of the deviations. The test is generally applicable (for reasons and for exceptions see 9, 11, 17, 41) only when there is a single linear restraint upon chi square, i.e., when the sum of the expected frequencies has been made to equal that of the observed, so that the number of degrees of freedom is one less than the number of deviations. The data are arranged in a table, similar to the one shown, with each column in the same relative position as the abscissa interval from which its data were taken. The chi square test is conducted in the usual way and its cumulative probability, P (X^2), is obtained. Then the total number of runs of plus and of minus deviations is counted among the deviations as they are arranged in the table. This number is referred to a probability table, supplied by David, which gives

$$\Pr(U \leqq U_o) = \sum_{U=2}^{U_o} \Pr(U \mid n_1 n_2), \quad \text{i.e., which gives the probability}$$

of the obtained number of runs cumulated from 2 to the obtained number and conditional upon the existence of n_1 plusses and n_2 minusses. (Since $\sum f_e$ has been made to equal $\sum f_o$, $\sum f_o - \sum f_e = 0$, i.e., the sum of the deviations must equal zero, and there must be at least one positive and one negative deviation. Therefore, since one run is impossible, the cumulation starts with two. However this qualification is automatically imposed whenever both n_1 and n_2 are different from zero, so any set of total-number-of runs tables is appropriate if entered with the obtained values of n_1 and n_2, neither of which can, in this application, equal zero.)

The chi square test and the toal number of runs test are independent. Therefore a single significance level can be obtained for the two tests by calculating their joint probability. This is somewhat complicated by the fact that chi square is a continuously distributed variable while the distribution of the total number of runs is discrete. However, David (9) has simplified matters by tabling this joint probability. Thus one obtains the product of $P(X^2)$, the cumulative probability of the obtained X^2, and P(U), the probability of the total number of runs cumulated from U = 2 to the obtained value. David's tables give the values of this product which are significant at the .05 and .01 levels of significance for values of $n_1 + n_2 \leqq 14$. It is particularly important that expected cell frequencies should be large enough for the binomial sampling distribution of "observed" frequencies to be well approximated by a normal distribution. This is the case because "an assumption implicit in the test would appear to be that for each X^2 cell there is an equal chance of obtaining a positive or a negative deviation". Furthermore, the independence of the chi square and run tests relates to the theoretical, continuously distributed chi square distribution, not to chi square as calculated from the sample. The discrepancy between the two "chi squares" is neglibible when expected cell frequencies are large, and effective independence can be expected to obtain; however, there is no certainty that the chi square and run tests continue to be independent when expected frequencies are small.

9. Extensions of Run Theory

Runs discussed so far have involved only two kinds of elements arranged in a linear sequence. However various probability formulae have also been derived for runs of like elements when there are more than two kinds of elements (34, 43, 49) and for runs where adjacency among elements can occur along two or more dimensions (3, 16, 24, 25, 26, 27, 31, 36, 37, 45). Such multiple-category and polydimensional runs are generally analysed on the basis of large sample theory, using critical ratios, rather than exact probabilities, since the exact distribution of such runs rapidly becomes difficult to tabulate as sample size increases.

BIBLIOGRAPHY

*T 1. BATEMAN, G., On the power function of the longest run as a test for randomness in a sequence of alternatives. *Biometrika*, 1948, 35, 97-112.

2. Baticle, E., Sur la probabilité des iterations dans le schema de Bernoulli *Comptes Rendus* (Paris), 1951, 472-473. Vol. 232.

3. Bose, R. C., On a problem of two dimensional probability. *Sankhyā*, 1950, 10, 13-28.

4. Bush, R. R., and Mosteller, F., *Stochastic models for learning*, New York: Wiley, 1955, 100-104.

5. Casanova, T., The use of the method of runs for testing the randomness of the order of examination items. *Journal of Experimental Education*, 1944, 12, 165-168.

6. Clark, A. L., Experimental probability. *Canadian Journal of Research*, 1934, 11, 658-664.

7. Cochran, W. G., An extension of Gold's method of examining the apparent persistence of one type of weather. *Quarterly Journal of the Royal Meteorological Society*, 1938, 64, 631-634.

8. Cochran, W. G., Statistical analysis of field counts of diseased plants. *Journal of the Royal Statistical Society (B)*, 1936, 3, 49-67.

TT* 9. David, Florence N., A χ^2 'smooth' test for goodness of fit. *Biometrika*, 1947, 34, 299-310.

10. David, Florence N., A power function for tests of randomness in a sequence of alternatives. *Biometrika*, 1947, 34, 335-339.

11. David, Florence N., Correlations between χ^2 cells. *Biometrika*, 1948, 35, 418-422.

T* 12. Dixon, W. J., A criterion for testing the hypothesis that two samples are from the same population. Annals of Mathematical Statistics, 1940, 11, 199-204.

13. Dixon, W. J., Power under normality of several non-parametric tests. Annals of Mathematical Statistics, 1954, 25, 610-614.

14. Epstein, B., Comparison of some non-parametric tests against normal alternatives with an application to life testing. Journal of the American Statistical Association, 1955, 50, 894-900.

15. Feller, W., An introduction to probability theory and its applications, Vol. I, New York: Wiley, 1950, 51-59, 264-278.

16. Finney, D. J., The significance of association in a square point lattice. Journal of the Royal Statistical Society, (B). 1947, 9, 99-103.

17. Fraser, D. A. S., Note on the χ^2 smooth test. Biometrika, 1950, 37, 447-448.

18. Gage, R., Contents of Tippett's "Random sampling numbers". Journal of the American Statistical Association, 1943, 38, 223-227.

19. Gold, E., Note on the frequency of occurrence of sequences in a series of events of two types. Quarterly Journal of the Royal Meteorological Society, 1929, 55, 307-309.

TT 20. Grant, D. A., Additional tables of the probability of "runs" of correct responses in learning and problem-solving. Psychological Bulletin, 1947, 44, 276-279.

T* 21. Grant, D. A., New statistical criteria for learning and problem solution in experiments involving repeated trials. Psychological Bulletin, 1946, 43, 272-282.

T 22. HOEL, P. G., Introduction to mathematical statistics, New York: Wiley, 1947, 177-182.

T 23. Hoel, P. G., Introduction to mathematical statistics, 2nd Ed., New York: Wiley, 1954, 293-299.

24. Krishna Iyer, P. V., Random association of points on a lattice. Nature, 1947, 160, 714.

25. Krishna Iyer, P. V., Random association of points on a lattice. Nature, 1948, 162, 333.

26. Krishna Iyer, P. V., The first and second moments of some probability distributions arising from points on a lattice and their applications. Biometrika, 1949, 36, 135-141.

27. Krishna Iyer, P. V., The theory of probability distributions of points on a lattice. Annals of Mathematical Statistics, 1950, 21, 198-217.

28. Kuznets, S., Random events and cyclical oscillations. Journal of the American Statistical Association, 1929, 24, 258-275.

29. Lehmann, E. L., Consistency and unbiasedness of certain nonparametric tests. Annals of Mathematical Statistics, 1951, 22, 165-179.

30. Lehmann, E. L., The power of rank tests. Annals of Mathematical Statistics, 1953, 24, 23-43.

31. Levene, H., A test of randomness intwo dimensions. (Abstract), Annals of Mathematical Statistics, 1946, 17, 500.

* 32. Mathen, K. K., A criterion for testing whether two samples have come from the same population without assuming the nature of the population. Sankyā, 1946, 7, 329.

33. Mood, A. M., On the asymptotic efficiency of certain non-parametric two-sample tests. Annals of Mathematical Statistics, 1954, 25,. 514-522.

34. MOOD, A. M., The distribution theory of runs. Annals of Mathematical Statistics, 1940, 11, 367-392.

35. Moore, P. G., A test for randomness in a sequence of two alternatives involving a 2 X 2 table. Biometrika, 1949, 36, 305-316.

36. Moran, P. A. P., Random associations on a lattice. Proceedings of the Cambridge Philosophical Society, 1947, 43, 321-328.

37. Moran, P.A.P., The interpretation of statistical maps. Journal of the Royal Statistical Society (B), 1948, 10, 243-251.

T* 38. Mosteller, F., Note on an application of runs to quality control charts. Annals of Mathematical Statistics, 1941, 12, 228-232.

39. Olmstead, P. S., Note on theoretical and observed distributions of repetitive occurrences. Annals of Mathematical Statistics, 1940, 11, 363-366.

T* 40. Ramachandran, G. and Ranganathan, J., A non-parametric two sample test. Journal of Madras University, Section B, 1953, 23, 76-91.

41. Seal, H. L., A note on the χ^2 smooth test. Biometrika, 1948, 35, 202.

* 42. STEVENS, W. L., Distributions of groups in a sequence of alternatives. Annals of Eugenics, 1939, 9, 10-17.

43. Sukhatme, B. V., On certain probability distributions arising from points on a line. Journal of the Royal Statistical Society, (B), 1951, 13, 219-232.

T 44. SWED, FRIEDA, S, and EISENHART, C., Tables for testing randomness of grouping in a sequence of alternatives. Annals of Mathematical Statistics, 1943, 14, 66-87.

45. Todd, H., A note on random association in a square point lattice. Journal of the Royal Statistical Society (B), 1940, 7, 78-82.

46, Uspensky, J. V., Introduction to mathematical probability, New York: McGraw-Hill, 1937, 77-84.

47. van der Waerden, B. L., Order tests for the two-sample problem II, and III. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1953, 56, 303-310, and 311-316.

* 48. WALD, A. and WOLFOWITZ, J., On a test whether two samples are from the same population. Annals of Mathematical Statistics, 1940, 11, 147-162.

49. WILKS, S. S., Mathematical Statistics, Princeton, N. J.: Princeton University Press, 1950, 200-207.

TT 50. Wilson, E. B., An introduction to scientific research, New York: McGraw-Hill, 1952, 266-268.

51. Wolfowitz, J., On the theory of runs with some applications to quality control. Annals of Mathematical Statistics, 1943, 14, 280-288.

52. Wolfowitz, J., Non-parametric statistical inference. In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Ed. by J. Neyman, Berkeley and Los Angeles: University of California Press, 1949, pp. 93-113.

CHAPTER IX

RUNS UP AND DOWN

A type of run test for trend can be obtained by defining a run as an unbroken sequence of increasing or decreasing observations. In this case the two kinds of events, "greater than the preceding observation" and "smaller than the preceding observation," are neither fixed in number nor of constant probability (since their probabilities depend on how "extreme" was the preceding observation). Thus the formulae developed in the preceding chapter are inappropriate. By investigating the probability for a given pattern of observation magnitudes, rather than a given pattern of dichotomized "events," the necessary formulae are obtained. Run tests of this type have used the total number of runs, the length of the longest run, or chi-square applied to frequencies of runs of various lengths, as the test statistic.

1. Introduction

Suppose that n observations have been taken on a continuously distributed variable and arranged in the order in which recorded. A continuously ascending sequence of observations will be defined as a run "up" and a monotonically decreasing sequence will be called a run "down". Now suppose that each observation is subtracted from the succeeding observation. There will be n-1 algebraic signs to replace the n original observations. A run "up" will now be more definitively indicated by a sequence of + signs, and a run "down" will be unambiguously identified by a run of - signs. The farther an observation is from the median of the series, the less likely it will be that the succeeding observation will depart from the median still farther. Therefore "plus" and "minus" are not constant probability events and probability formulae for runs up and down must be derived in the light of that fact.

Consider the probability that the i^{th} observation obtained initiates a run up of exactly S+1 observations so that the difference sign obtained by subtracting the i^{th} from the i+1st observation is the first + in a sequence of exactly S plusses. A run up of S+1 observations must begin with the first observation in the entire series when $n=S+1$and it must either begin with the first or end with the last observation when $n=S+2$. In order to examine the general case where the run can initiate, terminate or lie enclosed within the series, assume that $n \geqq S+3$. Consider first the probability that the series begins with a run up of exactly S+1 ascending observations. Let the first S+2 observations be replaced by their ranks, from 1 to S+2, in order of increasing magnitude. If the series is random, i.e., contains no true trend, each of the (S+2)! permutations of these S+2 observations is equally probable. But in order for the series to begin with a run up of exactly S+1 ascending observations, the S+2 ranks must be arranged so that: (a) the rank S+2, i.e., the highest among the S+2 observations, occupies the S+1st position, (b) any one of the remaining S+1 ranks occupies the S+2nd position, (c) the remaining S ranks are arranged in order of increasing size. Of these three requirements, (a) can be fulfilled in only one way, (b) can be accomplished in S+1 ways and (c) can then take place in only one way. So the probability that the series begins with a run of increasing observations of exactly length S+1 is $\frac{S+1}{(S+2)!}$. This is also the analogously derived probability that the series ends with a run up of exactly S+1 ascending observations, i. e., that a run up of S+1 observations begins with the n-S^{th} observation.

Now consider the probability that a run up of S+1 ascending observations begins at some specified position, i, where $2 \leqq i \leqq n - S - 1$, i. e., excluding the cases where the run begins or ends the series. Let the i-1st to the i+S+1st observations be ranked from 1 to S+3 in order of increasing magnitude. If the series is random, each of the (S+3)! permutations of order for these S+3 ranks is equally likely. But only in the following ways can the S+3 ranks be arranged so that the first is higher then the second, the second to the S+2nd form an ascending sequence, and the S+3rd is lower than the S+2nd: (a) Rank 1 occupies the 2nd position, rank S+3 occupies the next to last position, any one of the remaining S+1 ranks is placed in the first position, any one of the remaining S ranks is placed in the last position, and the remaining S-1 ranks are arranged in increasing order of magnitude from 3rd to second from last position. (b) Rank 1 occupies the second position, rank S+2 occupies the next to last position, rank S+3 occupies the first position, any one of the remaining S ranks is placed in the last position, and the remaining S-1 ranks are arranged in increasing order of magnitude from 3rd to second from last position. (c) Rank 2 occupies the second position, rank S+3 occupies the next to last position, rank 1 occupies the last position, any one of the remaining S ranks is placed in the first position, and the remaining S-1 ranks are arranged in increasing order of magnitude from 3rd to second from last position. (d) Rank 2 occupies the second position, rank S+2 occupies the next to last position, rank S+3 occupies the first position, rank 1 occupies the last position, and the remaining S-1 ranks are arranged in order of increasing magnitude from 3rd to second from last position. There is only one way in which a specified rank can be assigned to a specified position and only one way in which S-1 ranks can be arranged in order of increasing magnitude in S-1 positions. Therefore, the number of ways in which (a), (b), (c), and (d) can be accomplished is (S+1)S, S, S, and 1 respectively. The probability that a run up of exactly S+1 ascending observations begins at a predesignated position, i, when

$2 \leqq i \leqq n - s - 1$, is therefore $\frac{S^2+3S+1}{(S+3)!}$.

We have seen that when $n \geqq S + 3$, the probability that a run of exactly S+1 ascending observations begins with the i^th^ observation is $\frac{S+1}{(S+2)!}$ when i = 1 or when i = n-S and is $\frac{S^2+3S+1}{(S+3)!}$ when i is any one of the n-S-2 values between 2 and n-S-1. These are probabilities that the i^{th} observation initiates a run up of specified length, i.e., each probability is conditional upon the i-1st observation, if there is one, not being a continuation of the run. Otherwise viewed, each probability is conditional upon the i^{th} observation not being a continuation of any run up which began at some point earlier in the series. Therefore, since the probabilities do not refer to overlapping events, they can be summed over all possible values of i to obtain the expected number of runs of the specified type. Thus, when $n \geqq S+3$ the expected number of runs of ascending observations of length exactly S+1 or of plus difference signs of length exactly S is $\frac{2(S+1)}{(S+2)!} + \frac{(n-S-2)(S^2+3S+1)}{(S+3)!}$ which reduces to $\frac{n(S^2+3S+1) - (S^3+3S^2-S-4)}{(S+3)!}$. Following analogous derivations, it is clear that when n = S+2 the expected number of runs up of exactly S+1 observations is $\frac{2(S+1)}{(S+2)!}$ and when n = S+1 it is $\frac{1}{(S+1)!}$. (It should be noted that these derivations are based upon the n observations being in a random order, not upon each difference sign of a given type being equally likely, which is not the case.)

The expected number of runs up of ascending observations of length S+1 or longer is derived in a manner analogous to that already presented, dropping the restriction that the S+1st observation composing the run be followed by a lower observation. Thus assuming $n \geqq S+2$, one requires only that when i = 1 the S+1 observations beginning with the i^{th} are arranged in order of increasing magnitude and that when $2 \leqq i \leqq n - S$, in addition to the above requirement the i-1st observation is higher than the i^{th}. The expected number of runs of ascending observations of length S+1 or greater is therefore $\frac{1}{(S+1)!} + \frac{(n-S-1)(S+1)}{(S+2)!}$ or $\frac{n(S+1) - (S^2+S-1)}{(S+2)!}$ when $n \geqq S+2$ or $\frac{1}{(S+1)!}$ when $n = S+1$. And if 1 is substituted for S in the above formulas, the result is the expected number of runs of ascending observations of length S+1 = 2 or greater, or the number of runs of plusses of length 1 or greater. This expected number is $\frac{2n-1}{6}$ when $n \geqq S+2$ and 1/2 when n = S+1.

A run up and a run down commencing with the i^{th} observation are mutually exclusive events. Therefore to obtain the expected number of runs up <u>or</u> down, the expected number of runs up should be doubled. Variances for runs of either plusses or minuses of length S, or of length S or greater, have been given by Levene and Wolfowitz (7). The formulae for the general case, i.e., with S a variable, are lengthy. However, they are greatly shortened when S is given a specific value. For S = 1, $\sigma^2 = \frac{305\,n-347}{720}$, and for S = 2, $\sigma^2 = \frac{51,106\,n-73,859}{453,600}$. For $S \geqq 1$, $S \geqq 2$, and $S \geqq 3$, the respective variances are: $\frac{16\,n-29}{90}$, $\frac{57\,n-43}{720}$, and $\frac{21,496\,n-51,269}{453,600}$.

Consider the n observations ranked from 1 to n in order of increasing magnitude. There are n! permutations of these ranks, and the <u>expected number</u> of runs of a specified type is simply the total number of such <u>runs</u> which can be found in these n! permutations divided by the number of permutations, n!. On the other hand, the <u>probability</u> of at least one run of the type specified is the total number of <u>permutations</u> in which such a run can be found divided by the number of permutations, n!. Therefore the probability and expected number do not coincide when it is possible for more than one run of the specified type to be found in a single permutation. However, when $S > \frac{n-1}{2}$ the formulae already presented for the expected number of runs of a given variety also give the exact probability of occurrence for such runs. This appears to be the only situation, when dealing with runs up or down, in which an exact probability can be calculated without resort to a recursion formula.

2. Length of Longest Run Up or Down

Using a recursion formula Olmstead (9) has calculated and tabled exact probabilities for runs of like difference signs of length S or greater when $2 \leqq n \leqq 14$. For $n > 14$ Olmstead has tabled approximate probabilities calculated from asymptotic formulae (9, 13).

3. Total Number of Runs Up and Down

The total number of runs is simply the number of runs of plusses or minusses of length 1 or greater, and this was shown in Section 1, Introduction, to have an expected value of $\frac{2n-1}{3}$ and a variance of $\frac{16n-29}{90}$ when n is greater than 2. The total number of runs, r, is asymptotically normally distributed (6, 12), so for large values of n the significance of the total number of runs can be tested by treating r as a normal deviate and referring the critical ratio

$$\frac{r - \frac{2n-1}{3}}{\sqrt{\frac{16n-29}{90}}}$$

to normal tables. By reducing the absolute value of the numerator by 1/2, the critical ratio can be corrected for continuity.

If the total number of runs is r, then the series has reversed direction r-1 times, and a test based on the number of "turning points" is equivalent to one based on the total number of runs. The expected number of turning points, T, is $\frac{2n-4}{3}$ and its variance is the same as that for the total number of runs. Therefore the significance of the number of turning points can be tested by forming the critical ratio analogous to that given above, referring it to normal tables. When all tests concerned are applied to samples from normally distributed populations the turning point test has an asymptotic relative efficiency of zero with respect to the regression coefficient test and also with respect to each of eight distribution-free tests of randomness with which it was compared (10, 11). See Table I of Introduction.

4. Chi Square Applied to Run Frequencies

The expected number of runs of plusses or minusses of exactly length S was derived in Section 1, Introduction, and found to be $\frac{4(S+1)}{(S+2)!} + \frac{2(n-S-2)(S^2+3S+1)}{(S+3)!}$, and the expected total number of runs of plusses or minusses of all lengths was found to be $\frac{2n-1}{3}$, the former result requiring that $n \geqq S+3$ and the latter being contingent upon $n \geqq S+2$. However, if one regards the first and last runs as "incompleted" and counts only those runs which are preceded and followed by at least one run, the term $\frac{4(S+1)}{(S+2)!}$ in the first formula must be dropped since it represents the first and last runs, and the expected total number of runs must be reduced by 2. Thus the revised formulae become $\frac{2\,(n-S-2)\,(S^2+3S+1)}{(S+3)!}$ and $\frac{2\,n-7}{3}$, respectively. Substituting 1 and then 2 for S in the first revised formula, the expected number of runs of plusses or minusses of lengths 1 and 2 are found to be $\frac{5\,(n-3)}{12}$ and $\frac{11(n-4)}{60}$ respectively. Subtracting these two values from the expected total number of runs one obtains $\frac{4\,n-21}{60}$, the expected number of runs of plusses or minusses of length greater than 2.

Wallis and Moore (12, 8) have suggested a chi square test of significance applied in the usual way to the observed frequencies of "interior" runs of like signs of lengths 1, 2 and over 2, with the corresponding expected frequencies being $\frac{5\,(n-3)}{12}$, $\frac{11\,(n-4)}{60}$ and $\frac{4\,n\ -21}{60}$. There are 2 degrees of freedom one degree having been expended by obtaining n from the sample. The test, however, is an approximate one if the significance of the calculated chi square is obtained from the usual chi square tables. This is the case because the run lengths are not entirely independent of one another although the chi square test assumes that they are. Various empirically obtained "corrections" are offered by the authors for use when n exceeds 12. However, for $6 \leqq n \leqq 12$ they have provided a table of exact probabilities for the values of chi square as calculated from the sample. These were obtained by means of a recursion formula and give, in effect, that proportion of the n! permutations which yield a value of chi square as great or greater than the one tabled.

The test can be used as a test of randomness against either trend or correlation alternatives. In the latter application, if an x measurement and a y measurement have been taken on each of n objects, the objects are arranged in order of increasing magnitude of one continuously distributed variable and the run test is applied to measurements on the other variable. The authors point out, however, that "the conclusion occasionally depends upon which variate is chosen for arranging in order and which for counting the phase durations".

BIBLIOGRAPHY

1. Grant, A. M., Some properties of runs in smoothed random series. *Biometrika*, 1952, 39, 198-204.

2. Kendall, M. G., *The advanced theory of statistics*, Vol. II, London: Griffin, 1946, 124-127.

3. Kermack, W. O. and McKendrick, A. G., Some distributions associated with a randomly arranged set of numbers. *Proceedings of the Royal Society of Edinburgh*, 1936-37, 57, 332-376

* 4. KERMACK, W. O. and McKENDRICK, A. G., Tests for randomness in a series of numerical observations. *Proceedings of the Royal Society of Edinburgh*, 1936-37, 57, 228-240.

5. Krishna Iyer, P. V., Runs up and down on a lattice. *Nature* 1950, 166, 276.

6. Levene, H., On the power function of tests of randomness based on runs up and down. *Annals of Mathematical Statistics*, 1952, 23, 34-56.

7. LEVENE, H. and WOLFOWITZ, J., The covariance matrix of runs up and down. *Annals of Mathematical Statistics*, 1944, 15, 58-69.

8. Moore, G. H. and Wallis, W. A., Time series significance tests based on signs of differences. *Journal of the American Statistical Association*, 1943, 38, 153-164.

*T 9. OLMSTEAD, P. S., Distribution of sample arrangements for runs up and down. *Annals of Mathematical Statistics*, 1946, 17, 24-33; Bell Telephone System Technical Publications, Monograph 2289.

10. Stuart, A., Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. Journal of the American Statistical Association, 1954, 49, 147-157.

11. Stuart, A., The efficiencies of tests of randomness against normal regression. Journal of the American Statistical Association, 1956, 51, 285-287.

**T 12. Wallis, W. A. and Moore, G. H., A significance test for time series analysis. Journal of the American Statistical Association, 1941, 36, 401-409.

13.. Wolfowitz, J., Asymptotic distribution of runs up and down. Annals of Mathematical Statistics, 1944, 15, 163-172.

14. Wolfowitz, J., Note on runs of consecutive elements. Annals of Mathematical Statistics, 1944, 15, 97-98.

CHAPTER X

TESTS BASED ON EXTREME VALUES

The number of observations in a second sample which exceed (or which are exceeded by and therefore included within) observations of certain size rank in the first sample can be predicted if the samples are drawn from a common population, or can be used to test the hypothesis of a common population if the "second sample" has already been drawn. In either case, the probability is simply the proportion of all possible arbitrary reassignments of observations to samples in which the specified number of exceedances is found to occur. If certain assumptions can be made, the tests for identical populations become tests for location, dispersion or extreme reaction. An analogous but different mathematical approach permits the setting of tolerance limits.

1. Exceedances: Prediction

a. Rationale. Suppose that a sample of n observations has been taken from a continuously distributed, but otherwise unknown, population and it is desired to know the probability that N or more observations in some future sample of size m will exceed the rth observation, in order of increasing magnitude, in the already obtained sample. For convenience, let the first sample be designated X's, the second sample, Y's. Since the two samples are defined to be from the same population, the X's can be considered as a random sample of n observations "drawn" from the n+m observations comprising the two samples. Consider the sample of Y's to have been drawn and the n+m observations in the two samples to have been arranged in order of increasing magnitude, irrespective of sample, and labeled Z's with subscripts indicating rank:

$Z_1, Z_2, \ldots, Z_{r-1+m-N}, Z_{r+m-N}, \ldots, Z_{n+m-1}, Z_{m+n}$. Consider now the probability of drawing an "X sample" of n observations from these Z's so as to leave a remaining "Y sample" of m observations, N of which exceed the rth X in order of magnitude (and m-N of which are smaller than X_r). In order to obtain such a sample: (a) we must draw Z_{r+m-N} which becomes X_r, (b) we must draw any r-1 of the Z's smaller than Z_{r+m-N} of which there are r+m-N-1, and (c) we must draw any n-r of the Z's greater than Z_{r+m-N} of which there are m+n-(r+m-N) or n-r+N. There is only one way of doing (a), but there are $\binom{r+m-N-1}{r-1}$ ways of accomplishing (b) and $\binom{n-r+N}{n-r}$ ways of fulfilling requirement (c). Therefore there are $\binom{r+m-N-1}{r-1}\binom{n-r+N}{n-r}$ ways of performing the entire operation. Since there are $\binom{m+n}{n}$ ways of drawing the X sample, without these restrictions as to position, the probability of drawing an X sample which will leave N of the remaining observations greater than X_r is $\dfrac{\binom{r-1+m-N}{r-1}\binom{n-r+N}{n-r}}{\binom{n+m}{n}}$, and this is the probability that in a future sample of m observations from the same population, exactly N observations will exceed X_r. The probability of at least N exceedances is $$P_r\ (\text{Exceedances} \geq N) = \sum_{i=N}^{m} \frac{\binom{r-1+m-i}{r-1}\binom{n-r+i}{n-r}}{\binom{n+m}{n}}$$

b. Assumptions. Random sampling and no tied observations. The latter assumption is met, in theory, if the population is continuously distributed and measurements are precise.

c. Treatment of Ties. Tied observations, if their proportion is small, become a practical problem only if X_r is tied with other observations. Since Y observations are hypothetical, none will be tied with X_r. If X_r is tied with other X's, calculate the exceedance probability as many times as there are X's tied with X_r, each time letting r be a different one of the ranks the members of the tied group would have if not tied. For a conservative estimate, use the smallest or largest of these, whichever results in the greater conservatism, as the probability of N exceedances. If it is desired to minimize the error, use the average of the separately calculated probabilities.

d. Application. Obtain a sample of n observations from the population in question and rank them from smallest, 1, to largest, n. Letting subscripts indicate rank, the ordered observations will be: $X_1, X_2, \ldots, X_r, \ldots, X_{n-1}, X_n$. Treating ties as outlined above, the probability that of m future observations at least N of them will be larger than X_r, the magnitude of the obtained observation whose rank is r, is given by the last formula in "Rationale".

e. Discussion. The formula given for the probability of at least N exceedances over X_r, of course, also gives the probability that m-N or fewer future observations will be less than X_r. The formula applies to exceedances over the rth smallest observation or, since the rth rank from the bottom is the n-r+1th rank from the top, to the n-r+1th largest observation.

The point probability for exceedances can be evaluated by use of binomial tables. Let the exceedance probability formula,

$\dfrac{\binom{r-1+m-N}{r-1}\binom{n-r+N}{n-r}}{\binom{n+m}{n}}$, be multiplied by $\dfrac{p^n q^m}{p^n q^m}$ or, equivalently, by

$\dfrac{p^{r-1} q^{m-N} p^{n-r} q^{N} p}{p^n q^m}$. The formula thus becomes

$$\frac{p\left[\binom{r-1+m-N}{r-1} p^{r-1} q^{m-N}\right]\left[\binom{n-r+N}{n-r} p^{n-r} q^{N}\right]}{\left[\binom{n+m}{n} p^{n} q^{m}\right]} .$$

Each of the expressions in brackets is a binomial probability and can be read directly from binomial tables. The values p and q=1-p can be selected arbitrarily by the experimenter, so long as all p's are taken to be the same exactly tabled value. For convenience take p=q=1/2. Then P_r (Exceedances $\geq$ N) =

$$\sum_{i=N}^{m} \frac{(1/2)\binom{r-1+m-i}{r-1}(1/2)^{r-1+m-i}\binom{n-r+i}{n-r}(1/2)^{n-r+i}}{\binom{n+m}{n}(1/2)^{n+m}} =$$

$$(1/2)\sum_{i=N}^{m} \frac{\text{Point Bin. Pr.}(.50, r-1+m-i, r-1)\ \text{Point Bin. Pr.}(.50, n-r+i, n-r)}{\text{Point Bin. Pr.}(.50, n+m, n)}$$

f. Tables. Wilks (37) has published a short table of probabilities for exceedances over the smallest value of an obtained sample, i.e. r = 1. Epstein (4) has tabled exceedance probabilities for the case where the future sample is to be equal in size to the obtained sample, i.e. m = n. Rosenbaum (21) has tabled probabilities for exceedances over the largest value of an obtained sample, i.e. r = n. Gumbel and von Schelling (10) have graphed the probability of one or two exceedances over the largest or near-to-the-largest X value. Tables and graphs are also to be found in (11) and (13). Also see "Discussion" section for techniques of using binomial tables to obtain exceedance probabilities.

g. Sources. (4, 10, 11, 13, 21 , 36, 37)

2. Exceedances: Tests of Hypotheses

In the preceding section the probability was determined for at least N exceedances in a second, future, sample from the same population. If the second sample has actually been obtained, this same probability, derived in the same way, is the a priori probability for the obtained results under the null hypothesis that the two samples come from identical populations.

Exceedances therefore can be used to test the null hypothesis that two samples come from the same population under the assumption that the population is continuously distributed. In order to be able properly to use exceedance probability tables to test this hypothesis, X_r must be designated in advance of sampling, i. e. both the rank, r, and the sample, whether X or Y, determining the "reference point" for exceedances must be selected in advance.

Rosenbaum (21), Epstein (4), and Mathisen (13) have all suggested such tests. Rosenbaum uses exceedances over the largest X observation as his test statistic and has provided tables for it. Epstein uses exceedances over X_r, with r allowed to assume any preassigned value, but with the restriction that the two samples be of equal size. Tables are provided. Mathisen takes for X_r the median of an X sample containing an odd number of observations and provides a small table of probabilities for the number of observations in a second sample which will be lower than the median of the first. All three tests are, in effect, based on the premise that if the two samples are from identical populations the expected proportion of each sample above some arbitrarily designated value, X_r, should be the same. However, while identical populations insure that the proportion of each population above X_r is the same, the reverse is not true. The two populations can assume widely differing forms and, so long as their cumulative distribution functions are equal at the point X_r the null hypothesis will not be rejected more than α of the time. The above tests are therefore not consistent except for such classes of alternatives as slippage, i. e., f(y) = f(x+c) with c a constant (2).

If other X observations are tied with X_r, they should be treated as outlined in the preceding section. Y observations tied with X_r can be "assigned" positions. For a conservative test, count all Y observations tied with X_r as falling on whichever side of X_r which will be least conducive to rejection of the null hypothesis. To minimize error, half may be assigned above, half below X_r, an odd tied observation being treated "conservatively".

For a test at significance level α, reject if

$$\sum_{i=N}^{m} \frac{\binom{r-1+m-i}{r-1}\binom{n-r+i}{n-r}}{\binom{n+m}{n}} \leq \alpha$$

for a one-tailed test against the alternative hypothesis of excessive exceedances, i.e., that the proportion of values in the Y population which are greater than X_r exceeds the proportion of the X population which is greater than X_r. For a two-tailed test in which the alternative hypothesis is "either too many or too few exceedances", reject the null hypothesis if either the above summation or the summation taken from i=0 to i=N is less than $\frac{\alpha}{2}$.

This type of significance test is particularly useful when experimentation is costly in terms of time or material. All m of the Y observations need not necessarily be taken, since the null hypothesis can be rejected whenever the number of exceedances among the Y's reaches a certain value (determined by n, m, r and α). The test is especially appropriate for life testing since the experiment need last only long enough to identify X_r and for the number of exceedances to reach the rejection criterion (3).

3. Includances: Prediction

a. Rationale. Let the first sample from a continuously distributed but otherwise unknown population be arranged in order of ascending magnitude as follows: $X_1, X_2, \ldots, X_r, \ldots, X_s, \ldots, X_{n-1}, X_n$ (r and s being ranks which can be assigned any integral value from 1 to n so long as s is greater than r). The following derivation will obtain the probability that N observations, in a future sample of m observations, designated as Y's, will lie within the range of magnitudes whose endpoints are X_r and X_s. Consider the second sample to have been drawn and let the n+m observations be arranged in order of increasing magnitude, irrespective of sample, and labeled Z's with subscripts indicating rank: $Z_1, Z_2, \ldots, Z_{r+L}, \ldots, Z_{s+L+N}, \ldots Z_{n+m}$. The a priori probability that exactly L of the Y observations are smaller than X_r, N are between X_r and X_s, and m-L-N above X_s is the probability of drawing the X sample so as to consist of r-1 of the Z's below Z_{r+L}, Z_{r+L}, s-1-r of the Z's between Z_{r+L} and Z_{s+L+N}, Z_{s+L+N}, and n-s of the Z's above Z_{s+L+N}. This probability is

$$\frac{\binom{r-1+L}{r-1}\binom{1}{1}\binom{s-1-r+N}{s-1-r}\binom{1}{1}\binom{n+m-s-L-N}{n-s}}{\binom{n+m}{n}} .$$

This probability contains the restriction that exactly L of the m-N Y's outside of the X_r to X_s range shall fall below X_r. In order to remove this undesired restriction, the probability must be summed over all of the values from 0 to m-N which L can assume without changing N. The probability that exactly N of the Y's will fall between X_r and X_s is therefore

$$\sum_{L=0}^{m-N} \frac{\binom{r-1+L}{r-1}\binom{s-1-r+N}{s-1-r}\binom{n+m-s-L-N}{n-s}}{\binom{n+m}{n}}$$ and the probability that N or more Y's will fall between X_r and X_s is

$$\sum_{i=N}^{m}\ \sum_{L=0}^{m-i}\ \frac{\binom{r-1+L}{r-1}\binom{s-1-r+i}{s-1-r}\binom{n-s+m-L-i}{n-s}}{\binom{n+m}{n}}\ .$$

b. Assumptions. See 1, Exceedances: Prediction

c. Treatment of Ties. See 1. Observations tied with X_s should be treated separately but in the same way as observations tied with X_r.

d. Application. Last formula in "Rationale" gives required probability.

e. Discussion. The probability that N or more Y's will fall between X_r and X_s is, of course, also the probability that m-N or less of the Y's will fall outside the interval bounded by X_r and X_s.

This probability can be expressed in terms of several binomial probabilities. It becomes

$$\sum_{i=N}^{m}\ \sum_{L=0}^{m-i}\ \frac{p^2[\binom{r-1+L}{r-1}p^{r-1}q^{L}][\binom{s-1-r+i}{s-1-r}p^{s-1-r}q^{i}][\binom{n-s+m-L-i}{n-s}p^{n-s}q^{m-L-i}]}{[\binom{n+m}{n}p^{n}q^{m}]}$$

each of the bracketed expressions being obtainable from tables of the point binomial. The parameters p and q are chosen arbitrarily. For p = q = 1/2, the double summation becomes

$$\sum_{i=N}^{m}\ \sum_{L=0}^{m-i}\frac{(1/4)[\binom{r-1+L}{r-1}(1/2)^{r-1+L}]\,[\binom{s-1-r+i}{s-1-r}(1/2)^{s-1-r+i}]}{[\binom{n+m}{n}\,(1/2)^{n+m}]}$$

$$\cdot\ \frac{[\binom{n-s+m-L-i}{n-s}(1/2)^{n-s+m-L-i}]}{1}$$

Even with the help of binomial tables this probability is not quickly evaluated. By careful choice of the parameters n, m, r, s, N, L, the formula can be considerably simplified. Without such simplification the method will prove practical only when n and m are quite small or when tables of probabilities for N are available.

f. Tables. Wilks (37) has published a small table for the case where $r=1$, $s=n$. Rosenbaum (20) has produced an extensive table for the same case, but in terms of the probability for m-N Y values outside of the interval X_1 to X_n. Moses (14) has also published a small table for certain cases where $s=n-r+1$. Binomial tables can also assist in evaluating probabilities. See (e).

g. Sources. (14, 20, 36, 37)

4. Includances: Tests of Hypotheses

If the second sample has actually been obtained, the probability derived in the preceding section can be used to test the null hypothesis that the two samples are from the same continuously distributed population. The values n, m, r, s and α must, of course be selected in advance of sampling, which must be random. Rosenbaum (20) proposes includances as a test of equal dispersions for two populations known to have the same median. He has provided extensive probability tables for the number of Y's which fall outside of the interval whose endpoints are X_1 and X_n. If medians are not known to be equal, his test becomes a test for identical populations. Moses (14) uses includances to test the null hypothesis that an experimental and a control group belong to the same population against the alternative hypothesis that the treatment to which the experimental group is subjected tends to increase the scores of some individuals and reduce those of others ("defensive responses"). Moses takes as his test statistic the number of X's equal to or included between X_r and X_{n-r+1} plus the number of Y's included between these endpoints. Since the number of X's in this interval is predetermined, the probability for the obtained statistic is the same as the probability for the number of Y includances. A small table of probabilities is given.

X scores tied with X_r and X scores tied with X_s should be dealt with separetely but in the same way as outlined under 1. Exceedances: Prediction, for observations tied with X_r. Y scores tied with X_r and Y scores tied with X_s should, for a conservative test, all be counted as falling within or outside the X_r to X_s interval, whichever is least conducive to rejection of the null hypothesis. If it is desired to minimize the error, half of Y's tied with X_r should be counted as falling inside the interval, half outside, and likewise for Y's tied with X_s, odd tied Y's being dealt with conservatively.

For a one-tailed test of the null hypothesis of identical populations against the alternative of excessive includances, reject at the level α if

$$\sum_{i=N}^{m} \sum_{L=0}^{m-i} \frac{\binom{r-1+L}{r-1}\binom{s-1-r+i}{s-1-r}\binom{n-s+m-L-i}{n-s}}{\binom{n+m}{n}} \leq \alpha .$$

If the alternative is too few includances, reject at the level α if the double summation equals or exceeds $1 - \alpha$. For a two-tailed test, reject at the level α if the double summation $\leq \frac{\alpha}{2}$ or $\geq 1-\frac{\alpha}{2}$. The above formula is valid for the desired probability <u>only</u> if <u>previous to sampling</u> it is specified which sample is to be the X sample and which the Y sample. The values of r, s, n, m, and α must also be decided upon before the samples are obtained.

If r and s are taken to be 1 and n respectively so that the interval is that included between the smallest and largest X observations, the probability is greatly simplified. The first and last combinatorial expressions in the numerator become 1. Summing from L=0 to L=m-i, therefore, amounts simply to multiplying

$$\frac{\binom{s-1-r+i}{s-1-r}}{\binom{n+m}{n}} \text{ or } \frac{\binom{n-2+i}{n-2}}{\binom{n+m}{n}}$$

by m-i+1. The probability that N of m Y observations will fall within the endpoints of a sample of n X observations from the same continuously distributed population thus becomes

$$\sum_{i=N}^{m} \frac{(m-i+1)\binom{n-2+i}{n-2}}{\binom{n+m}{n}} \text{ or } \frac{m!\ n(n-1)}{(n+m)!} \sum_{i=N}^{m} \frac{(n-2+i)!\ (m-i+1)}{i!} .$$

5. A Univariate Tolerance Limit

a. Rationale. While confidence limits specify a region within which a population parameter is inferred to lie, tolerance limits enclose a region within which a specified proportion of the entire population is inferred to exist.

Let a sample of n observations be taken from a continuously distributed population, f(x), and arranged in order of increasing magnitude with subscripts indicating rank in that order. The proportion of the unknown parent population which is smaller than X_r, the r^{th} smallest sample observation, is $\int_0^{x_r} f(x)\ dx$ or $F(x_r)$, the small case x indicating the same value as X, but located in the population rather than the sample. $F(x_r)$ is therefore the probability, P, of a sample observation being less than x_r. The a priori probability that in a random sample of size n, r-1 observations will fall below x_r, one observation at x_r, and n-r observations above x_r is given by the multinomial law for partitions: $\frac{n!}{(r-1)!\ 1!\ (n-r)!} [F(x_r)]^{r-1}$

$\cdot [1-F(x_r)]^{n-r} [f(x_r)dx_r]$. Substituting P for $F(x_r)$, this becomes

$$\frac{n!}{(r-1)!\ (n-r)!} P^{r-1} (1-P)^{n-r}\ dP.$$

This states the probability that the r^{th} ordered sample observation occupies the area of the population distribution curve (i. e density function) whose ordinate is $f(x_r)$ and whose base is dx_r. Equivalently, it is the probability that exactly a proportion $P=F(x_r)$ of the parent population lies below x_r. By integrating from $P=\lambda$ to $P=1$ we obtain the probability that a proportion λ or more of the parent population lies below the r^{th} smallest sample observation.

Thus $\int_{\lambda}^{1} \frac{n!}{(r-1)!(n-r)!} P^{r-1}(1-P)^{n-r} dP$ gives the desired probability.

This can be evaluated by means of tables of the incomplete beta function since

$$1 - \int_{\lambda}^{1} \frac{n!}{(r-1)!(n-r)!} P^{r-1}(1-P)^{n-r}dP = \int_{0}^{\lambda} \frac{n!}{(r-1)!(n-r)!} P^{r-1}(1-P)^{n-r}dP$$

$$= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)} \int_{0}^{\lambda} P^{r-1} (1-P)^{n-r} dP = I_{\lambda} (r, n-r+1).$$

The probability sought is therefore $1-I_{\lambda} (r, n-r+1)$, or if tables of the incomplete beta function are not available, binomial tables can be used since $I_{\lambda} (r, n-r+1) = \sum_{i=r}^{n} \binom{n}{i} \lambda^{i} (1-\lambda)^{n-i}$.

By obvious symmetry the probability that a proportion λ of the population lies below the r^{th} smallest sample observation is also the probability that a proportion λ of the population lies above the r^{th} largest sample observation, i.e., the $n-r+1^{th}$ ordered observation.

b. Assumptions. Random sampling from a continuously distributed population. The latter assumption was implicitly introduced in the derivation when the probability of an observation above x_r was taken as one minus the probability of an observation below x_r. This leaves the probability for an observation equal to x_r to be zero which is the case only if f(x) is continuous in the region of x_r.

c. Treatment of Ties. Ties are problem only if they involve the r^{th} ordered sample value. In this case, if the proportion of tied observations is small, one of the following treatments may be employed. Take a new r, r' which refers to the middle ordered observation in the tied group to which the old x_r belonged, and calculate λ using r' and $x_{r'}$,instead of r and x_r. Alternatively, calculate λ for each of the ordered observations tied with x_r and either use the average λ , or the most "conservative" λ .

d. Application. Decide upon the values to be used for r and n prior to sampling. Then take a sample of n observations from the population in question, arrange them in order of increasing magnitude and select x_r the r^th ordered sample value. If it is desired that the tolerance level is to be $1-\alpha$ that a proportion λ or more of the parent population lies below x_r, solve

$$\int_{\lambda}^{1} \frac{n!}{(r-1)!\,(n-r)!} P^{r-1} (1-P)^{n-r}\, dP \geq 1-\alpha \text{ for } \lambda.$$ This can be accomplished simply by referring to tables of the incomplete beta function or to tables of the cumulative binomial (See Rationale). Actually, if any three of the values n, r, λ, α are preselected, the fourth can be found by solving the above formula.

e. Discussion. There is an element of inaccuracy in this method of obtaining tolerance limits. The derivation is based on the formula

$$\frac{n!}{(r-1)!\,1!\,(n-r)!} [\,F(x_r)]^{r-1} [\,1-F(x_r)\,]^{n-r} f(x_r)\, dx_r$$

in which the "event", one observation in the region dx_r, is (a) given a probability of occurrence, $f(x_r)\, dx_r$, which must be zero since the probabilities, $F(x_r)$ and $1-F(x_r)$, for the other two multinomial categories together equal 1, (b) is regarded as having occurred once in n trials. The occurrence in a finite number of trials of a predesignated event with zero probability is, of course, implausible. The ambiguity, and inexactitude, result from the mixture, in the same formula, of terms implying a discrete distribution, i.e. the multinomial, with terms relating only to a continuous distribution, i.e., $f(x_r)\, dx_r$. The net result is inaccuracy in the order of dx_r, or, in more practical terms, the distance between successive ordered observations, namely x_r and x_{r+1}. The error therefore should be between zero and $x_{r+1}-x_r$. Since a sample of n observations randomly divides its population distribution into n+1 intervals each of which, on the average, contains a proportion $\frac{1}{n+1}$ of the population, the error in λ would not be expected to exceed $\frac{1}{n+1}$. See the section on confidence limits for quantiles for a similar discussion.

f. Tables. The required probabilities can be obtained from tables of the incomplete beta function (17, 26), by special use of tables of the cumulative binomial (25) or, for the case where $r=1$, directly from a small table prepared by Wilks (37).

g. Sources. (11, 15, 17, 23, 24, 25, 26, 36, 37)

6. Univariate Tolerance Limits

Let a sample of n observations, capital X's, be drawn from a continuously distributed but otherwise unknown population f(x) and arranged in order of increasing magnitude $X_1, X_2, \ldots, X_r, \ldots, X_s, \ldots, X_{n-1}, X_n$. These n ordered observations divide the unknown population from which they came into n+1 intervals: $-\infty$ to x_1, x_1 to x_2, ..., x_{r-1} to x_r, ..., x_{s-1} to x_s, ..., x_{n-2} to x_{n-1}, x_{n-1} to x_n, x_n to $+\infty$, small case x's denoting the same magnitudes as the large case X's, but magnitudes located in the parent population, not the obtained sample.

The probability of drawing an observation smaller than some value x_i is simply $F(x_i)$, the cumulated probability for values of x less than x_i. This $F(x_i)$ is known to have a uniform distribution from 0 to 1, so that its probability is the same for every x_i, i.e., is independent of i, (See Mood (I pp. 107-108) for proof). And since the probability for $F(x_i)$ is independent of i, the probability for $F(x_i) - F(x_{i-1})$ is independent of i. However, this is the propprtion of the parent population within the interval x_{i-1} to x_i. Therefore the proportion of the parent population to be enclosed between successive ordered sample observations is independent of the rank of the observations.

Stated slightly differently, each of the n+1 intervals has exactly the same probability of enclosing any given proportion of the parent population. In the last section the probability that a proportion λ or more of the parent population lies below X_r was found. Since there are r intervals below X_r, the derived probability is also the probability that a proportion λ or more of the population lies in any preselected r intervals between successive ordered sample values. It is therefore the probability that λ or more of the population lies between X_i and X_{i+r}, if the values i and r are selected prior to sampling.

For assumptions, application, etc., see the preceding section. Tied observations are a problem if they include either X_i or X_{i+r}. If there is one such group of ties, they should be dealt with as indicated in the preceding section. If X_i and X_{i+r} are both members of tied groups, each group should be treated separately, but in a fashion analogous to that outlined previously. For sources, see (1, 15, 17, 22, 23, 24, 25, 26, 35, 36, 37).

7. Multivariate Tolerance Limits

Ingenious methods of setting tolerance limits for multivariate distributions have been discussed by Wald (34), and others (5-9, 32, 33). For the bivariate case Wald selects four integers a, b, c, d before sampling n observations from a continuously distributed bivariate population. After obtaining the sample, he discards the a observations with the smallest, and the b observations with the largest, x values; then, of the remaining n-a-b observations, he discards the c observations with the smallest, and the d observations with the largest, y values. The tolerance region is the rectangle bounded by the a^{th} smallest and the b^{th} largest X and by the c^{th} smallest and the d^{th} largest of the n-a-b Y's between the X boundaries. Tukey (32, 33) has generalized the method of "cuts" by which the tolerance region is obtained and has extended the applicability of the method to discontinuously distributed populations. Fraser (5, 6, 7) has further developed the method so that instead of a predetermined method of making cuts, each cut can be made in a manner determined by the outcome of previous cuts. For details of application, see the referenced articles.

BIBLIOGRAPHY

1. Birnbaum, Z. W. and Zuckerman, H. S., A graphical determination of sample sizes for Wilks' tolerance limits. *Annals of Mathematical Statistics*, 1949, 20, 313-316.

2. Bowker, A. H., Note on consistency of a proposed test for the problem of two samples. *Annals of Mathematical Statistics*, 1944, 15, 98-101.

3. Epstein, B., Comparison of some non-parametric tests against normal alternatives with an application to life testing. *Journal of the American Statistical Association,* 1955, 50, 894-900.

*T 4. Epstein, B., Tables for the distribution of the number of exceedances. *Annals of Mathematical Statistics*, 1954, 25, 762-768.

5. Fraser, D. A. S., *Nonparametric methods in statistics,* New York: Wiley, 1957.

6. Fraser, D. A. S., Nonparametric tolerance regions. *Annals of Mathematical Statistics*, 1953, 24, 44-45.

7. Fraser, D. A. S., Sequentially determined statistically equivalent blocks. *Annals of Mathematical Statistics*, 1951, 22, 372-381.

8. Fraser, D. A. S. and Guttman, I., Tolerance regions, *Annals of Mathematical Statistics*, 1956, 27, 162-179.

9. Fraser, D. A. S. and Wormleighton, R., Nonparametric estimation IV., *Annals of Mathematical Statistics*, 1951, 22, 294-298.

T 10. GUMBEL, E. J. and von SCHELLING, H., The distribution of the number of exceedances. *Annals of Mathematical Statistics,* 1950, 21, 247-262.

T 11. Harris, L. B., On a limiting case for the distribution of exceedances, with an application to life testing. *Annals of Mathematical Statistics*, 1952, 23, 295-298.

12. Kemperman, J. H. B., Generalized tolerance limits. *Annals of Mathematical Statistics*, 1956, 27, 180-186.

*T 13. Mathisen, H. C., A method of testing the hypothesis that two samples are from the same population. *Annals of Mathematical Statistics*, 1943, 14, 188-194.

*T 14. Moses, L. E., A two-sample test. *Psychometrika*, 1952, 17, 239-247.

T 15. MURPHY, R. B., Non-parametric tolerance limits. *Annals of Mathematical Statistics*, 1948, 19, 581-589.

16. Noether, G. E., On a connection between confidence and tolerance intervals. *Annals of Mathematical Statistics*, 1951, 22, 603-604.

T 17. Pearson, K., *Tables of the incomplete beta-function.* London: Office of Biometrika, University College, 1934.

18. Pillai, K. C. S., A note on ordered samples. *Sankhyā*, 1948, 8, 375-380.

19. Robbins, H., On distribution-free tolerance limits in random sampling. *Annals of Mathematical Statistics*, 1944, 15, 214-216.

*T 20. Rosenbaum, S., Tables for a nonparametric test of dispersion. *Annals of Mathematical Statistics*, 1953, 24, 663-668.

*T 21. Rosenbaum, S., Tables for a nonparametric test of location. *Annals of Mathematical Statistics*, 1954, 25, 146-150.

22. Scheffé, H. and Tukey, J. W., A formula for sample sizes for population tolerance limits. *Annals of Mathematical Statistics*, 1944, 15, 217.

23. Scheffé, H. and Tukey, J., Non-parametric estimation. I. Validation of order statistics. Annals of Mathematical Statistics, 1945, 16, 187-192.

24. Shewhart, W. A., Contribution of statistics to the science of engineering. New York: Bell Telephone System Monograph B-1319, 1941

T 25. Staff of the Computation Laboratory, Tables of the cumulative binomial probability distribution. Cambridge, Mass.: Harvard University Press, 1955, pp. xliii-xlvi.

T 26. THOMPSON, CATHERINE M., Tables of percentage points of the incomplete beta-function. Biometrika, 1941, 32, 151-181.

27. Thompson, W. R., Assumption economy in scientific statistical analysis, Mimeographed from Division of Laboratories and Research, N. Y. Department of Health, Albany, N. Y.

* 28. Thompson, W. R., Biological applications of normal range and associated significance tests in ignorance of original distribution forms. Annals of Mathematical Statistics, 1938, 9, 281-287.

* 29. Thompson, W. R., On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. Annals of Mathematical Statistics, 1936, 7, 122-128.

30. Thompson, W. R., Statistical methods for evaluation of diagnostic and other procedures. I An objective weeding-out process applicable to material used in surveys of diagnostic variability. Human Biology, 1949, 21, 17-34.

31. Thompson, W. R. et al, Excerpts from Annual Reports of the Division of Laboratories and Research, N. Y. State Department of Health, Albany, N. Y.: 1948, pp. 32-33, 1949, pp. 27-29, 1953, pp. 26-27.

32. Tukey, J. W., Non-parametric estimation II. Statistically equivalent blocks and tolerance regions - the continuous case. Annals of Mathematical Statistics, 1947, 18, 529-539.

33. Tukey, J. W., Nonparametric estimation, III. Statistically equivalent blocks and multivariate tolerance regions - the discontinuous case. Annals of Mathematical Statistics, 1948, 19, 30-39.

34. Wald, A., An extension of Wilks' method for setting tolerance limits. Annals of Mathematical Statistics, 1943, 14, 45-55.

* 35. WILKS, S. S., Determination of sample sizes for setting tolerance limits. Annals of Mathematical Statistics, 1941, 12, 91-96.

36. WILKS, S. S., Order Statistics. Bulletin of the American Mathematical Society, 1948, 54, 6-50.

*T 37. WILKS, S. S., Statistical prediction with special reference to the problem of tolerance limits. Annals of Mathematical Statistics, 1942, 13, 400-409.

CHAPTER XI

TESTS BASED ON THE MAXIMUM DEVIATION BETWEEN TWO CUMULATIVE DISTRIBUTIONS

If the cumulative distribution for an obtained sample and either (a) the cumulative distribution of the population from which it was drawn, or (b) the cumulative distribution for a second sample from the same population, are plotted on the same graph, the maximum deviation between the two cumulatives will be independent of the form of the sampled population. Its probability fraction can be obtained, however; therefore the maximum deviation can be made the test statistic for distribution-free tests of goodness of fit or tests of whether two samples were drawn from identical populations. By confining the test to the lower portion of a sample cumulative, the test can be made especially efficient for life testing. Tables of probabilities for the maximum deviation can be used to set confidence bands for the cumulative distribution of the sampled population.

1. Maximum Deviation Tests for Goodness of Fit to an Hypothesized Population

a. Rationale. Let F(x) be the true population cumulative distribution of x and let F(x) be plotted as ordinate against x as abscissa. Now suppose that a sample of n observations is drawn from the x population and that the sample cumulative distribution Sn(x) has been plotted on the same graph with F(x). Thus Sn(x) is a step function which rises in steps of 1/n or multiples thereof. Let d be the maximum ordinatewise deviation between the smooth curve F(x) and the step function Sn(x). It has been proven (23, 43) that the probability of d taking any specified value is independent of the form of F(x) so long as F(x) is continuously distributed. This can be seen as follows. The probability that an observation drawn from the x population will be below some value x_i is simply $F(x_i)$, the value of the cumulative distribution at the point x_i. The probability that exactly r observations in a sample of n observations will lie below x_i is $\binom{n}{r} [F(x_i)]^r [1 - F(x_i)]^{n-r}$. And if this occurs, a proportion, r/n, of the sample has fallen below x_i, and this is the ordinate of the sample cumulative step function, Sn(x), at the abscissa x_i. Therefore $\binom{n}{r} [F(x_i)]^r [1 - F(x_i)]^{n-r}$gives the probability that the difference in ordinates between the population cumulative distribution and the sample cumulative step function will be $F(x_i) - r/n$ at the abscissa point x_i. Let $F(x_i) - r/n = c$.

Then $F(x_i) = \frac{r}{n} + c$ and the a priori probability that $F(x_i)-Sn(x_i)=c$ is $\binom{n}{r} [\frac{r}{n} + c]^r [1-\frac{r}{n} - c]^{n-r}$. The latter expression depends only upon c, n and r of which the former is a constant specified in the probability statement and the latter two are parameters of the sample, not of the population. Therefore the probability that $F(x_i) - Sn(x_i) = c$ is independent of F(x), i.e., is independent of the form of the distribution of the parent population. This is obviously true for any value of c, and since x_i was chosen arbitrarily it is also true for any value of x. Thus the probability that the maximum absolute deviation equals or exceeds d, i.e., $Pr\ (\max |Sn(x) - F(x)| \geq d)$, is independent of the form of F(x) so long as F(x) is continuously distributed.

Therefore, if the probability of d can be derived by assuming that x has a uniform distribution, the result can be generalized to any continuous distribution. Following this approach, let x have a uniform distribution with range from 0 to 1. Then F(x) = x, and F(x) is a line of constant slope rising from an ordinate of zero to an ordinate of 1. Now divide the populatioan range of xs into n equal abscissa intervals. Since F(x) is a line of constant slope, each of the n equal abscissa intervals contains the same proportion, 1/n, of the population. Let n_1, n_2, ... n_n be the obtained number of sample values falling in the first, second ... n^th^ interval. The expected proportion of sample values falling in any given interval is, of course, 1/n. Therefore the a priori probability of the obtained results is given by the multinomial and is $\frac{n!}{n_1!\,n_2!\ldots n_n!}\left(\frac{1}{n}\right)^{n_1}\left(\frac{1}{n}\right)^{n_2}\ldots\left(\frac{1}{n}\right)^{n_n}$ or $\frac{n!}{n_1!\,n_2!\,\ldots\,n_n!}\left(\frac{1}{n}\right)^{n}$. This is the probability of a specified pattern of interval-occupancy, n_1, n_2, ... n_n. Corresponding to each pattern of interval-occupancy is a pattern or set of ordinate differences at interval end points: At the end of the first interval the ordinate of F(x) is 1/n and that of Sn(x) is n_1/n; at the end of the second interval the ordinate of F(x) is 2/n and that of Sn(x) is $\frac{n_1+n_2}{n}$, etc. The probability of the pattern of interval-occupancy is therefore equally the probability of the set of n-1 ordinate differences. Therefore by examining all possible patterns of interval-occupancy, selecting those for which the corresponding set of ordinate differences contains an ordinate difference of d or greater, and summing the probabilities, $\frac{n!}{n_1!\,n_2!\,\ldots\,n_n!}\left(\frac{1}{n}\right)^{n}$, associated with these critical d's, one obtains the probability that at one of the abscissa points, 1/n, 2/n, ... i/n, ... n/n, the ordinate difference between F(x) and Sn(x) will equal or exceed d. See Figure 3.

P_R (ANY GIVEN SET OF ORDINATES FOR $S_n(x)$ AT INTERVAL ENDPOINTS) = P_R (CORRESPONDING SET OF n_i)

THE LATTER IS $\frac{n!}{n_1!\,n_2!\dots n_n!}\left(\frac{1}{n}\right)^n$. THUS, $P_R\{S_n(x) > F(x) + d \text{ AT AN ENDPOINT}\} = \sum \frac{n!}{n_1!\,n_2!\dots n_n!}\left(\frac{1}{n}\right)^n$ TAKEN OVER ALL SETS OF n_i IN WHICH $S_n(x) > F(x) + d$ AT AN ENDPOINT.

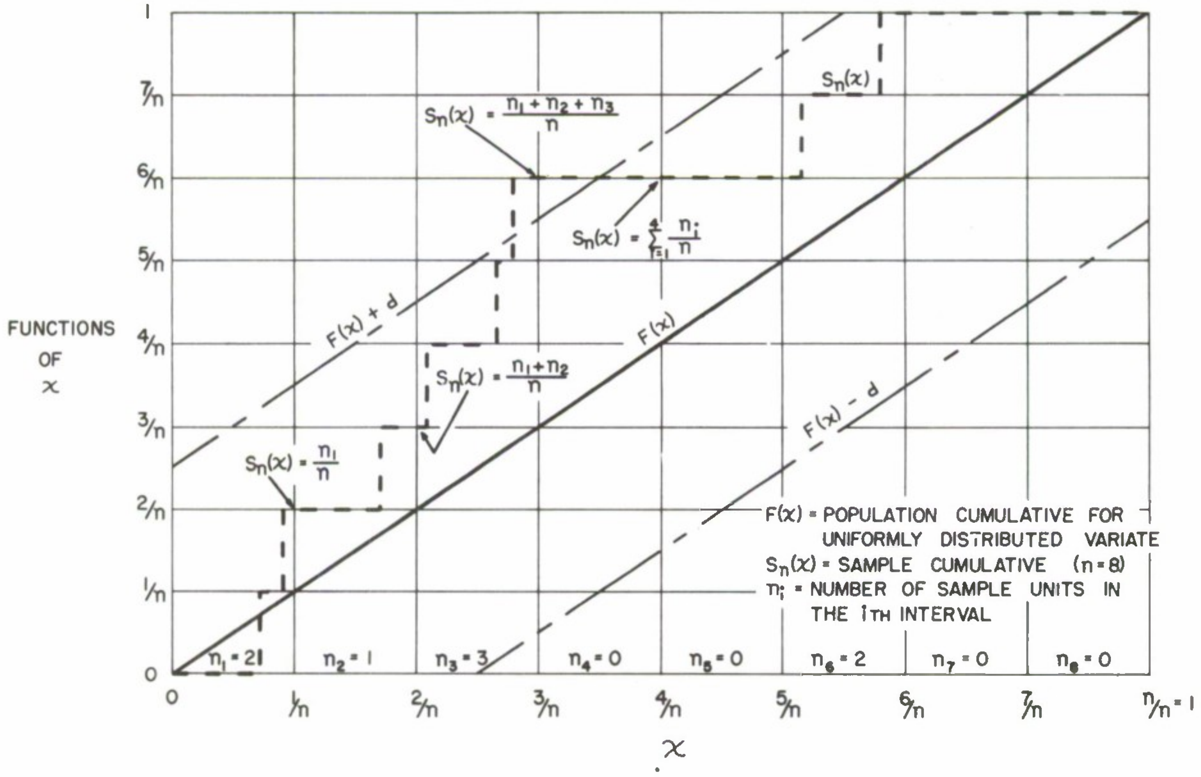


Figure 3.

$P_R\{S_n(x) > F(x) + d$ WITHIN THE iTH INTERVAL, GIVEN THAT $S_n(x) < F(x) + d$ AT THE INTERVAL ENDPOINTS$\} = \mathcal{A}_i^{\,n_i}$

WHERE $\mathcal{A}_i$ = PROPORTION OF INTERVAL WIDTH OUTSIDE OF $F(x) + d$ AT THE HIGHEST ORDINATE TAKEN BY $S_n(x)$ WITHIN THE iTH INTERVAL AND n_i = NUMBER OF SAMPLE UNITS IN THE iTH INTERVAL

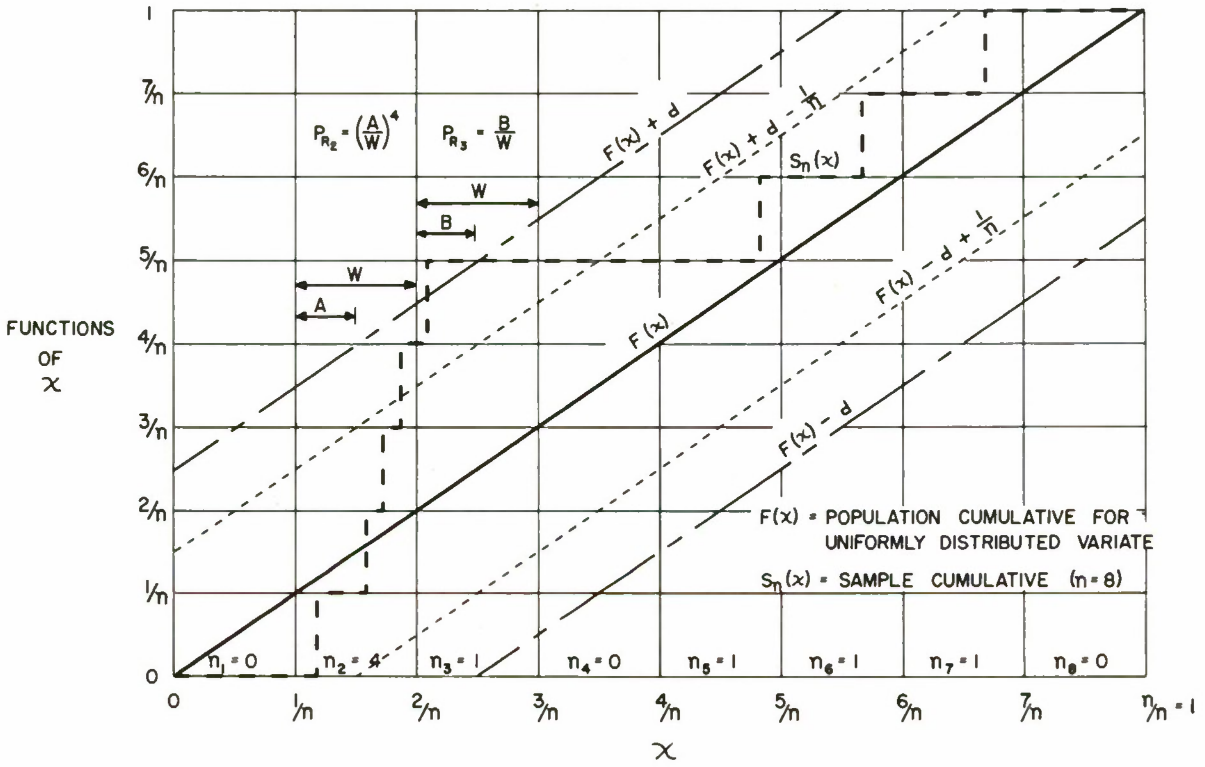


Figure 4.

However, a maximum absolute deviation of d or greater can occur within an interval without also occurring at its beginning or end. F(x) rises an ordinate distance of 1/n from the beginning to the end of an interval. Therefore, if $S_n(x)$ is greater than $F(x)+d-\frac{1}{n}$ but less than $F(x)+d$, at the upper endpoint of an interval, $S_n(x)$ may exceed $F(x)+d$ within the interval. Whether it does so or not depends upon the abscissa at which $S_n(x)$ rises to its greatest ordinate within the interval. (The greatest ordinate is 1/n greater than the next-to-greatest ordinate.) In order for $S_n(x)$ to exceed $F(x)+d$, it must assume its maximum ordinate before $F(x)+d$ exceeds that ordinate. Let a horizontal line be drawn across the width of the interval at the highest ordinate taken by $S_n(x)$ within the interval, and let K be that part of the horizontal line which lies outside of the confidence band, $F(x)+d$. In order for $S_n(x)$ to exceed $F(x)+d$, its maximum ordinate must overlap with K, and for this to happen all n_i must have abscissae beneath K. If p_i is the proportion of the interval width represented by K, then the probability that a randomly selected one of the n_i will lie below K is p_i, and the probability that all n_i units will lie below K is $p_i^{(n_i)}$. This is the probability that when $F(x)+d-\frac{1}{n}<S_n(x)<F(x)+d$ at the upper endpoint of the i^th interval, $S_n(x)> F(x)+d$ within the interval. See Figure 4. Similarly, if $F(x)-d<S_n(x)<F(x)-d+\frac{1}{n}$ at the lower endpoint of the i^th interval, let a horizontal line be drawn across the width of the interval at the lowest within-interval ordinate of $S_n(x)$, let L be that part of the line lying below $F(x)-d$, and let p'_i be the proportion of the interval width represented by L. The probability that $S_n(x)<F(x)-d$ within the interval is $p'_i{}^{(n_i)}$ and since K and L cannot have any abscissae in common, $p_i^{(n_i)}$ and $p'_i{}^{(n_i)}$ are mutually exclusive probabilities and can be added. Let Q_i be the sum of these two probabilities. Now consider those so far uncounted patterns of interval occupancy at all of whose endpoints $|F(x)-S_n(x)|<d$ and at some of whose endpoints $|F(x)-S_n(x)|>d-\frac{1}{n}$. The probability of each such pattern is $\frac{n!}{n_1!\, n_2! \ldots n_n!} (\frac{1}{n})^n$ as before. The probability that, given such a pattern of interval-occupancy, $| F(x) - S_n(x) | \geqq d$ within an interval depends not only upon the values of the Q_i, but upon the number of nonzero Q_i. If there is only one nonzero Q_i, then the probability $| F(x) - S_n(x) | \geqq d$ will be that Q_i times $\frac{n!}{n_1! n_2! \ldots n_n!} (\frac{1}{n})^n$.

If there are two nonzero Q_i, Q_1 and Q_2 say, then

$$P_r (| F(x) - S_n(x) \quad | \geqq d) = (Q_1 + Q_2 - Q_1 Q_2) \frac{n!}{n_1!\, n_2! \ldots n_n!} (\frac{1}{n})^n$$

since we are interested only in whether or not $S_n(x)$ exceeds the confidence bands, not in how often it does so within a given pattern of interval-occupancy. The probabilities are then summed over all critical patterns, i.e., those in which the maximum absolute deviation at the endpoints of one of the n intervals lies between $d - 1/n$ and d. This sum plus the previously obtained probability that d will be equalled or exceeded at an endpoint is the probability that max $| S_n(x) - F(x) | \geqq$ d at any mutual abscissa value.

Probabilities obtained in this manner are appropriate for a two-sided test. Probabilities for a one-sided test can be derived in analogous fashion by substituting "maximum deviation in a single predesignated direction", d', for "maximum absolute deviation", d, in the above. Thus instead of Pr (max $| Sn(x) - F(x) | \geqq d$) one obtains either Pr (max $(Sn(x) - F(x)) \geqq d'$) or Pr (max $(F(x) - Sn(x)) \geqq d'$).

b. Null Hypothesis. The parent population from which the sample was drawn is identical to, i.e., is completely and exactly defined by, the hypothesized population whose cumulative distribution is F(x).

c. Assumptions. Sampling is random, observations are independent, and the sampled population is continuously distributed.

d. Treatment of Ties. Although ties cause the test to become imprecise, they require no special modification of procedure. So long as the proportion of tied observations is small, the tabled probabilities will probably be very close approximations to the true ones.

When n is so large that tables whose probabilities are derived from asymptotic formulae must be used, ties cause the probability error to be in the conservative direction. If the true probability that max $| Sn(x) - F(x) | \geqq d$ is $\propto$, the tabled probability will be no smaller than $\propto$ so rejection will occur less frequently than would be the case if there were no ties. And if the true probability that max $| Sn(x) - F(x) | < d$ is $1 - \propto$, the tabled probability will be no greater than $1 - \propto$ and confidence limits obtained from the tables at the nominal $1 - \propto$ level of confidence will have a true confidence level equalling or exceeding that level (12, 22).

e. Efficiency. Van der Waerden (42) compared the power of the unidirectional maximum deviation test (at a significance level of .01) with that of the one-sided most powerful parametric test when both the sampled and the hypothesized populations were normally distributed with variance of 1, differing only in location. The unidirectional maximum deviation test was less powerful than the classical test with the power discrepancy increasing as sample size increased from 2 to 3 to 5. At n = 5 its efficiency had dropped to about .65. Massey (34) compared the smallest maximum absolute deviations detectable with probability .50 by the d test and by the chi-square test for $\propto$s of .05 and .01 and ns ranging from 200 to 2000. The d test was found to be superior to chi square in all of the 46 cases examined.

Massey (30, 31) has found the maximum absolute deviation test to be consistent provided that the sampled population is continuously distributed, but biassed for finite n. He has also obtained a lower bound for its power. Birnbaum (5) has found bounds for the power of the one-sided, i.e., maximum unidirectional deviation, test.

f. Application. Plot the cumulative distribution of the hypothesized population and the cumulative distribution, i.e. step function, of the sample on the same graph as shown in Figure 3 . Find the maximum ordinatewise deviation, d, between the two cumulative distributions. Enter the probability tables with d and n to determine the significance of the result.

g. Discussion. Much of the literature on maximum absolute deviation methods relates to the setting of confidence bands for an hypothesized population. Thus if Pr (max $| Sn(x) - F(x) | \geqq d) = \propto$, there will be a confidence level of 1 - α that Sn(x) will stay entirely within the band between two curves whose ordinates are F(x) + d and F(x) - d. Or if Pr (max $\{Sn(x) - F(x)\} \geq d'$) = α there is a probability of 1 - α that Sn(x) will never reach or exceed F(x) + d'. It is to be noted that d' at the level α is not identical to d at the level 2 α although when n $\leqq$ 100 and $\alpha \leqq$ .05 they are approximately equal (35).

The derivation given under "Rationale" was chosen for its conceptual simplicity. The method outlined is not the most efficient means of obtaining probabilities. Probabilities for the maximum absolute deviation have generally been obtained by means of recursion formulae. However, probabilities for the maximum unidirectional deviation can be obtained by use of a single exact formula derived by Birnbaum and Tingey (7).

The relative merits of chi square and the maximum absolute deviation test have been discussed by a number of authors. (4, 20, 34). The d test is superior to chi square in the following ways. The d test requires only the assumption of a continuously distributed population (other than the usual assumptions of randomness and independence) while chi-square requires, among others, the assumption that observed frequencies are normally distributed about their expected frequencies; thus the d test is distribution-free for all sample sizes while chi-square becomes distribtuion-free only when an infinite-sized sample permits the normality assumption to be fulfilled. The exact distribution of d is known and tabled for small sample sizes, while the exact distribution of chi-square is known and tabled only for infinite sized samples. The d' test can be used to test for deviations in a given direction, i.e., can be used as a one-sided test, while chi-square cannot. The d test uses ungrouped data, every observation representing a point at which the "goodness of fit" is examined; chi-square loses this information by requiring that data be grouped into cells. Furthermore by using ungrouped data the d test avoids the hazards and pitfalls associated with choice of interval size and selection of starting point in chi-square tests of fit and no correction for continuity is required by the d test. The d test can be applied to data which become available sequentially from smallest to largest, computations being continued only up to the point at which rejection occurs; it thus has an "efficiency" aspect not present in chi-square. Confidence bands can be easily established on the basis of the distribution of d, while chi-square has no such analogous property. More is known about the power of the d test than of chi square, and the information presently available suggests that in general it is the more powerful test. Chi square, on the other hand, is superior to d in the following ways. Chi square does not require that the hypothesized population be completely known in advance of sampling. Certain population parameters can be estimated from the sample and the resulting degree of "artificial" fit between obtained sample and hypothesized population can be taken account of and prevented from biassing the probability of significance by making the appropriate reduction in degrees of freedom. No such adjustment is possible with the d test, which requires that the hypothesized population be completely known and specified a priori. Chi square can be partitioned and added, very useful properties which the d statistic does not possess. Finally, chi square can be applied to discrete populations. The d test, however, is not incapable of such applications. When the assumption of continuity is not met, the probability of d is expressed by an inequality rather than an equation. The result is that the true probability that $d \geqq h$ (or that $d' \geqq h'$) is no greater than the tabled probability. Therefore in tests of significance the true probability of rejection may be smaller, but not greater, than the nominal probability, α. And in setting confidence limits, the true probability of inclusion within the limits may be greater, but not smaller, than the nominal probability of inclusion, $1 - \alpha$. In both cases the probability error is a "conservative" one. See (12, 20, 22).

h. Tables. Critical values of d at standard significance levels have been tabled by Miller (35) for all values of n from 1 to 100, approximate formulae having been used. A smaller table has been published by Massey (34). Probabilities that d will be less than c/n have been tabled by Birnbaum (4) for all values of n from 1 to 100, and Massey (29) has published a less extensive table. The limiting distribution of d or its equivalent has been given by a number of authors. Massey gives the values of d required for significance at standard significance levels when n is infinite (34) and the probability, at $n = \infty$, that $d < \lambda / \sqrt{n}$ for various values of λ (29). The latter probability has been tabled, in terms of λ rather than d, by Kolmogorov (22, 23). The limiting distribution of λ has been tabled by Smirnov (39).

Critical values of the maximum unidirectional deviation between sample step function and population cumulative distribution have been tabled by Miller (35) for all values of n from 1 to 100, standard significance levels being used. His probabilities are based on asymptotic formulae when n exceeds 20. A smaller table of such values has been published by Birnbaum and Tingey(7).

i. Sources. 3-12, 14-15, 17, 19-23, 26, 29-31, 34-37, 39, 42-44.

2. Related Tests of Fit

A statistic somewhat similar to that outlined in 1, Maximum Deviation Tests for Goodness of Fit to an Hypothesized Population, has been considered by a number of writers. It is

$n\omega_n^2 = n\int_{-\infty}^{\infty} (Sn(x) - F(x))^2 \, dF(x)$ which can be equivalently expressed as $n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} [\frac{2i-1}{2n} - F(x_i)]^2$. The statistic $n\omega_n^2$ is distribution free, and requires only the assumptions of random and independent sampling from a continuously distributed population. Its probabilities have been tabled for samples of size 1, 2 and 3 (28) and of size n = infinity (1, 28).

Anderson and Darling (1, 2) have proposed a modification of the above statistic which involves the application of a weight function to $(Sn(x) - F(x))^2$. They have also proposed (1) to modify the maximum absolute deviation test, described in 1, Maximum Deviation Tests for Goodness of Fit to an Hypothesized Population, by applying a weight function to $|Sn(x) - F(x)|$. These and related tests are discussed in (1, 2, 3, 11, 28, 38).

3. Truncated Maximum Deviation Tests for Identical Populations.

a. Rationale. Suppose that two samples, one of n observations labeled xs and the other of m observations labeled ys, have been drawn from continuously distributed populations and that the experimenter wishes to test whether or not the sampled populations are identical. Let Sn(x) and Sm(y) be the cumulative step functions of the x and y samples respectively, and let them be plotted on the same graph. Finally let d_r be the maximum difference in ordinates between the two step functions at any abscissa value less than or equal to the r^{th} x observation in order of ascending size. The probability that d_r equals or exceeds some predesignated value, h, has been tabled and can be used to test the hypothesis of identical populations.

Let the x observations be arranged in order of increasing size, $x_1, x_2, \ldots, x_i, \ldots, x_r, \ldots, x_n$ and let the number of y observations smaller than x_1 be designated m_1, the number of y observations between x_1 and x_2 be m_2, etc., so that m_i is the number of sample ys between x_{i-1} and x_i, and let the number of y observations greater than x_r be represented by M. The a priori probability for any set of such frequencies, $m_1, m_2, \ldots m_r$, M, is $\binom{M+n-r}{M}/\binom{m+n}{m}$. This can be proved as follows: If the two samples are from the same population, the sample designations x and y are arbitrary. The set $m_1, m_2, \ldots, m_r$, M may then be regarded as having been obtained by drawing labels, without replacement, from a population consisting of n x labels and m y labels and applying them, in the order drawn, to the m + n observations arranged in order of increasing size. The first m_1 labels must be ys, the next must be an x, then m_2 ys in succession followed by another x, etc. The probability of drawing the required label on any given draw is, of course, the remaining number of labels of the required type divided by the remaining number of labels of both types. Thus the denominator of the probability fraction is m + n on the first draw, $m+n-1$ on the second, and 1 on the last, and the product of these denominators is simply $(m+n)!$. The numerators will be

$$\{(m)(m-1)(m-2)\ldots(m-m_1+1)\}\ \{n\}\ \{(m-m_1)(m-m_1-1)\ldots$$

$$(m-m_1-m_2+1)\}\ \{(n-1)\}\ \ldots$$ etc., or, more concisely

$$\frac{m!}{(m-m_1)!}\ n\ \frac{(m-m_1)!}{(m-m_1-m_2)!}\ (n-1)\ \frac{(m-m_1-m_2)!}{(m-m_1-m_2-m_3)!}\ (n-2)\ \ldots$$

$\dfrac{(m_r+M)!}{M!}\ (n-r+1)(n-r+M)!$, which, after making the obvious cancellations, reduces to $\dfrac{m!}{M!}\ \dfrac{n!}{(n-r)!}\ (n-r+M)!$ or $\binom{n-r+M}{M}\ m!\ n!$. Dividing this numerator by the denominator $(m+n)!$, the resulting probability fraction is $\dfrac{\binom{n-r+M}{M}}{\binom{m+n}{m}}$. This is the probability that if observations from a sample of n xs and m ys are arranged in order of increasing size m_1 ys will be less than x_1, m_2 ys will lie between x_1 and x_2, etc., m_r ys will lie below x_r and M ys will lie above x_r. Obviously it is also the probability that m_1 ys will lie below x_1, m_1+m_2 ys will lie below x_2 etc., etc., and $m_1+m_2+\ldots+m_r$ ys will lie below x_r. And therefore it is the probability that at the abscissae x_1, x_2, x_3, ..., x_r the ordinates of the step function Sm(y) will be m_1/m, $\dfrac{m_1+m_2}{m}$, $\dfrac{m_1+m_2,+m_3}{m}$, ...

$\dfrac{m_1+m_2+m_3+\ldots+m_r}{m}$. At the same abscissae, the ordinates of Sn(x) jump a distance $1/n$, remaining constant at abscissae values in between. So the above probability is also the probability that at abscissae infinitesimally smaller than x_1, x_2, ..., x_r, the difference in ordinates between the two step functions will be

$\frac{o}{n} - \frac{m_1}{m}$, $\frac{1}{n} - \frac{m_1+m_2}{m}$; . . . , $\frac{r-1}{n} - \frac{m_1+m_2+ \ldots +m_r}{m}$ and

that at abscissae infinitesimally larger than $x_1, x_2, \ldots, x_{r-1}$ the differences in ordinates will be $\frac{1}{n} - \frac{m_1}{m}$, $\frac{2}{n} - \frac{m_1+m_2}{m}$, . . .

$\frac{r-1}{n} - \frac{m_1+m_2+ \ldots m_{r-1}}{m}$. The maximum absolute deviation, d_r, must be an ordinate difference in one of these two sets since, in the interval between x_i and x_{i+1}, the ordinate of Sn(x) remains constant while the ordinate of Sm(y) has its lowest value at x_i and its highest value at x_{i+1}. Thus the pattern of xs and ys, when arranged in order of increasing size, determines the maximum ordinatewise difference between the x and y sample step functions, and the probability that d_r equals or exceeds h is simply the sum of the probabilities of the arrangements of xs and ys in which $d_r \geqq h$. Otherwise stated, if $d_r \geqq h$ in K of the $\binom{m+n}{m}$ distinguishable arrangements of xs and ys, then the a priori probability that $d_r \geqq h$ is $K \binom{n-r+M}{M} / \binom{m+n}{m}$. If m and n are both very small, K can be determined by forming all patterns of xs and ys and counting the patterns for which $d_r \geqq h$. For larger values of m and n, recursion formulae are used to determine K.

The ordinate of Sn(x) reaches a height of r/n when an abscissa of x_r is reached. Therefore, if h is selected to be a value greater than r/n, then at any abscissa up to and including x_r, the ordinate of Sn(x) cannot exceed that of Sm(y) by a difference of h or more, although the reverse may occur. Thus when h > r/n, the test is one-sided in the sense that the null hypothesis can be rejected only because of an excess of ys over xs in the region below x_r.

Even when $h < r/n$, the number of xs below x_r is limited while the number of ys in this region is not; therefore, other things being equal, a large d_r is more likely to be the result of an excess of ys over xs in this region than of the reverse. The result is that the test is more likely to reject when there are too many ys below x_r than when there are too few, the bias increasing as r decreases.

This situation can be remedied and the test made unbiassedly two-sided by taking the maximum ordinatewise deviation below the r^{th} x or the r^{th} y, in ascending order, whichever is the larger. This, however, requires some modifications in derivations and formulae. Let d'_r be the maximum absolute deviation below x_r or y_r, whichever is larger. If $x_r > y_r$, then at least r ys lie below x_r, or, otherwise stated, M can be no greater than m-r. Thus the probability that $d'_r \geqq h$ and that $x_r > y_r$ is obtained by taking as K that number of arrangements in which $d_r \geqq h$ counted only from those arrangements in which $x_r > y_r$ or, equivalently, in which $M \leqq m-r$. Identifying the modified K as K',

$\Pr (d'_r \geqq h,\ x_r > y_r) = K' \binom{n-r+M}{M} / \binom{m+n}{m}$. Likewise,

$\Pr (d'_r \geqq h,\ y_r > x_r) = K'' \binom{m-r+N}{N} / \binom{m+n}{m}$ with K'' and N defined analogously to K' and M. Since $x_r > y_r$ and $y_r > x_r$ are mutually exclusive events, the probability that $d'_r \geqq h$, when d'_r is the maximum ordinatewise deviation occurring below whichever of the two values x_r and y_r is the larger, is simply the sum of the separate probabilities for these mutually exclusive events. Thus $\Pr (d'_r \geqq h) = \dfrac{K' \binom{n-r+M}{M} + K'' \binom{m-r+N}{N}}{\binom{m+n}{m}}$.

Some of these probabilities have also been tabled.

b. Null Hypothesis. Each of the $\binom{m+n}{m}$ distinguishable arrangements of xs and ys is equally likely to be the pattern obtained when the sample observations are arranged in order of increasing size. The null hypothesis will be true if the two samples come from the same population. It will be false, but will be rejected at the same level, $\propto$, as if it were true, if the two samples have been drawn from populations which are identical at values less than or equal to the critical value, x_r or y_r, and nonidentical at values above it.

c. Assumptions. Observations are drawn randomly and independently from continuously distributed populations.

d. Treatment of Ties. A relatively small number of ties are a practical problem only if an x and a y observation are tied for the abscissa value at which the maximum ordinatewise deviation, d_r, occurs. In this case, for a conservative test, the x and y should be slightly separated so as to give their ordinates the lesser deviation, and the maximum deviation, d_r, should be redetermined by examination of the entire graph. Or, to minimize error, break such ties in all possible ways, find d_r for each such way and obtain its probability, then use the average of these probabilities.

e. Efficiency. Epstein (18) empirically tested the relative efficiencies of the Wilcoxon test for unpaired observations, the "fully two-sided" version of the present test (in which d'_r is chosen from below max x_r, y_r), Epstein's version of the exceedances test, and the Wald-Wolfowitz total number of runs test. The tests were applied to two hundred pairs of samples of ten observations from each of two populations differing in means but having normal distributions and equal variances. The order in which the tests are listed above is the order of their efficiency, from best to worst, in detecting the difference between the population means.

f. Application. Forty type x and forty type y light bulbs are placed on life test. It is decided in advance to reject the hypothesis of identical life-expectancy populations if, by the time the fifth bulb of each type has blown, an ordinatewise deviation of probability $\leqq .05$ has occurred. Therefore $m=n=40$, $r = 5$, d'_r is the maximum ordinatewise deviation below x_5 or y_5, whichever is larger, and $\propto = .05$. The bulbs blow in the following order

$x_1, x_2, x_3, y_1, x_4, x_5, x_6, x_7, y_2, x_8, x_9, y_3, x_{10}, x_{11}, x_{12}.$

The test is halted when x_{12} blows because $12-3 = 9$ and Tsao's tables (40) show that for $m = n = 40$ and $r = 5$ a $d'_5 \leqq 8/40$ has probability .96, so a $d'_5 \geqq 9/40$ is significant at less than the .05 level.

g. Discussion. In respect to its derivation the present test is closely related to tests based upon exceedances.

h. Tables. Tsao (40), who originated the test, has tabled the probability that $d_r \leqq c/m$ for certain equal sized samples between $m = n = 3$ and $m = n = 40$ with r never exceeding 10. He has also prepared (40) a similar table for the probability that $d'_r \leqq c/m$.

i. Sources. (18, 40).

4. Maximum Deviation Tests for Identical Populations

a. Rationale. Suppose that r is set equal to n in the preceding test. The test statistic becomes d_n, the maximum difference in ordinates between Sn(x) and Sm(y) at any abscissa value below x_n. But at the abscissa x_n, the ordinate of Sn(x) is 1. The deviation between the two step functions cannot be greater at abscissae above x_n than it is at x_n. So the criterion d_n is equivalent to using as test statistic d, the maximum ordinatewise deviation between Sn(x) and Sm(y) at any common abscissa. In the preceding section the probability that, at some abscissa value less than x_r, max $|$ Sn(x) - Sm(y) $| \geqq h$ was found to be $K \binom{M+n-r}{M} / \binom{m+n}{m}$ where K was the number of distinguishable arrangements of xs and ys resulting in a $d_r \geqq h$. Substituting n for r, the probability becomes $K \binom{M+n-n}{M} / \binom{m+n}{m}$ which reduces to $K/\binom{m+n}{m}$. This was to be expected since $\binom{M+n-r}{M}$ is the number of arrangements of n xs and m ys in which m_1 ys are below x_1, m_2 ys are between x_1 and x_2, ..., m_r ys are between x_{r-1} and x_r, and M ys are above x_r.

When $r < n$, there are a number of ways in which M ys can be located above x_r, each distinguishable pattern of arrangement of M ys and $n - r$ xs constituting a different way. When $r = n$, there is only one way in which the M ys can be located above x_r. The distribution of the ys among the xs is completely specified, so the specification can be met by only one of the $\binom{m+n}{m}$ distinguishable patterns of arrangement of xs and ys. The maximum absolute deviation test, d, is therefore a special case of the truncated maximum absolute deviation test, d_r, described in the preceding section. The test can, of course, be made one-sided by substituting d', the maximum unidirectional deviation, for d and K', the number of the $\binom{m+n}{m}$ arrangements, in which the maximum unidirectional deviation equals or exceeds a specified value, h', for K. Thus, $Pr\ (\max \{Sn(x) - Sm(y)\} \geqq h') = K'/\binom{m+n}{m}$ for a one-sided test; and for a two-sided test $Pr\ (\max |Sn(x) - Sm(y)| \geqq h) = K/\binom{m+n}{m}$.

b. Null Hypothesis. Each of the $\binom{m+n}{m}$ distinguishable arrangements of xs and ys is equally likely to be the pattern obtained when the sample xs and ys are arranged in order of increasing size. This will be the case if the two samples are drawn from the same population.

c. Assumptions. See 3, Truncated Maximum Deviation Tests for Identical Populations.

d. Treatment of Ties. See 3. When m and n are both so large that tables whose probabilities are derived from asymptotic formulae must be used, ties cause the probability error to be in the conservative direction. The tabled probability will be no smaller than the true probability, so rejection will occur less frequently than would be the case if there were no ties (12, 22).

e. Efficiency. Applied to samples of size 5 and infinity from normal populations with equal variances but different means, the maximum absolute deviation test has an efficiency of .65 relative to Student's t-test, for both one-sided and two-sided tests. In the same situation, but with samples of sizes 5 and 6 the test is more efficient than the total number of runs test, but less efficient than the

Mann-Whitney test or the X-test (41). Applied to equal-sized samples of size 3, 4, or 5 from normal populations with equal variances and different means, the maximum absolute deviation test was more efficient than Westenberg's median test and less efficient than the Mann-Whitney test (13). It is more efficient than the total number of runs test and less efficient than the Mann-Whitney test when applied to large samples against the nonparametric alternatives investigated by Lehmann (25). (See Introduction).

The test has been proved consistent by Massey (30) provided only that the sampled populations are continuously distributed. See also (24). However, the test is biassed for finite n (24, 30, 31).

f. Application. Let the sample data be represented by the following table, observations being listed in increasing order of size.

x-observation	Corresponding ordinate of Sn(x)	y-observation	Corresponding ordinate of Sm(y)	Difference in ordinates
-512	1/16			1/16
-509	2/16			2/16
-487	3/16			3/16
		-422	1/4	1/16
		-415	2/4	5/16
-409	4/16			4/16
		-398	3/4	8/16
		-360	4/4	12/16
-341	5/16			11/16
-312	6/16			10/16
-275	7/16			9/16
-202	8/16			8/16
-111	9/16			7/16
-58	10/16			6/16
-14	11/16			5/16
9	12/16			4/16
21	13/16			3/16
75	14/16			2/16
156	15/16			1/16
201	16/16			0/16

The maximum ordinatewise deviation is 12/16 which for samples of sizes 4 and 16 is found, by using Massey's tables (32) to have a probability of .034 of being equalled or exceeded. Therefore the hypothesis of a common population would be rejected if a significance level of .05 were used.

g. Discussion. Drion (16) has derived exact probabilities, without resort to recursion formulae, by use of random walk methods. His formulae, however, require samples to be of equal size. He finds the probability that $d \geqq c/n$ to be $\dfrac{2\left[\binom{2n}{n-c} - \binom{2n}{n-2c} + \binom{2n}{n-3c} - \binom{2n}{n-4c} + \ldots\right]}{\binom{2n}{n}}$,

"the series being continued as long as $n - kc \geqq 0$". The sample sizes are, of course, n and m = n. The probability that the maximum unidirectional deviation, Sn(x) - Sn(y), exceeds c/n is found to be

$$\binom{2n}{n-c} \Big/ \binom{2n}{n}.$$

Drion has also used random walk to investigate the probability that, disregarding the endpoints whose ordinates are zero and one, two sample step functions will not intersect, i.e., that either one of the sample step functions will lie entirely above the other. If the two samples are of equal size and come from the same population this probability is $\frac{1}{2n-1}$. If the samples come from the same population but are of different sizes, n and m, and if n and m are coprime, the probability is $\frac{2}{n+m}$.

h. Tables. Massey has tabled the probability that d will not exceed specified values for equal sized samples with $1 \leqq m = n \leqq 40$ (33), for equal or unequal sized samples with $m \leqq 10$ and $n \leqq 10$, and for samples of selected larger sizes (32). A small table for use with equal-sized samples has been published by Drion (16). The limiting cumulative distribution of $d\sqrt{\frac{mn}{m+n}}$ has been tabled by Smirnov (39), thereby permitting the approximate probability of a given d to be obtained when m and n are both very large. Goodman (20) has published a table of probabilities for the maximum unidirectional deviation between ordinates of step functions of equal-sized samples for sample sizes ranging from 1 to 50.

i. Sources. 8-16, 19-22, 24-25, 30-33, 36, 39, 41.

5. A Large Sample Test Using Grouped Data

Marshall (27) has proposed an approximate test, for use when m and n are large, which involves grouping the data into class intervals. The range of the variables is divided into j + 1 intervals by the selection of j arbitrary points, and the unidirectional difference in step function ordinates is measured at each of these points. These differences are then summed. The sum, S, is normally distributed in the limit with mean zero and variance

$$\left(\frac{1}{m} + \frac{1}{n}\right)\left(\sum_{i=1}^{j} P_iQ_i + 2\sum_{i=1}^{j-1}\sum_{k=i-1}^{j} P_iQ_k\right).$$

The values of P_i, however, must be obtained from their maximum likelihood estimates,

$$\hat{P}_i = \frac{n\ Sn(i) + m\ Sm(i)}{m+n}$$

where Sn(i) and Sm(i) are the ordinates of the two step functions at the abscissa point i which is one of the j arbitrarily chosen points dividing the data into intervals. The test is conducted by referring the critical ratio to normal tables. Its asymptotic power efficiency is .64 for j = 1, .91 for j = 5, and .94 for j = 10 when used to test for a difference in means between two normally distributed populations with equal variances.

BIBLIOGRAPHY

**T 1. Anderson, T. W. and Darling, D. A., Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 1952, 23, 193-212.

T 2. Anderson, T. W. and Darling, D. A., A test of goodness of fit. *Journal of the American Statistical Association,* 1954, 49, 765-769.

3. Birnbaum, Z. W., Distribution-free tests of fit for continuous distribution functions. *Annals of Mathematical Statistics*, 1953, 24, 1-8.

TT 4. Birnbaum, Z. W., Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association*, 1952, 47, 425-441.

5. Birnbaum, Z. W., On the power of a one-sided test of fit for continuous probability functions. *Annals of Mathematical Statistics,* 1953, 24, 484-489.

6. Birnbaum, Z. W. and Pyke, R., On some distributions related to the statistic D_n. *Annals of Mathematical Statistics*, 1958, 29, 179-187.

*TT 7. BIRNBAUM, Z. W. and TINGEY, F. H., One-sided confidence contours for probability distribution functions. *Annals of Mathematical Statistics*, 1951, 22, 592-596.

8. Blackman, J., An extension of the Kolmogorcv distributions. *Annals of Mathematical Statistics*, 1956, 27, 513-520.

9. Blackman, J., Correction to "An extension of the Kolmogorov distribution." *Annals of Mathematical Statistics*, 1958, 29, 318-324.

10. Chung, K. L., An estimate concerning the Kolmogoroff limit distribution. *Transactions of the American Mathematical Society.* 1949, 67, 36-50.

11. Darling, D. A., The Kolmogorov-Smirnov, Cramér-von Mises tests. *Annals of Mathematical Statistics*, 1957, 28, 823-838.

12. David, H. T., *Discrete populations and the Kolmogoroff-Smirnov tests*, Report No. SRC-21103D27, Statistical Research Center, University of Chicago, Nov. 1952.

13. Dixon, W. J., Power under normality of several non-parametric tests. *Annals of Mathematical Statistics*, 1954, 25, 610-614.

14. Donsker, M. D., Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics*, 1952, 23, 277-281.

15. Doob, J. L., Heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics*, 1949, 20, 393-403.

T 16. Drion, E. F., Some distribution-free tests for the difference between two empirical cumulative distribution functions. *Annals of Mathematical Statistics*, 1952, 23, 563-574.

17. Dwass, M., On several statistics related to empirical distribution functions. *Annals of Mathematical Statistics*, 1958, 29, 188-191.

18. Epstein, B., Comparison of some non-parametric tests against normal alternatives with an application to life testing. *Journal of the American Statistical Association*, 1955, 50, 894-900.

19. Feller, W., On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics*, 1948, 19, 177-189.

20. Goodman, L. A., Kolmogorov-Smirnov tests for psychological research. *Psychological Bulletin*, 1954, 51, 160-168.

21. Kemperman, J. H. B., Some exact formulae for the Kolmogorov-Smirnov distributions. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1957, 60, 535-540.

T 22. Kolmogoroff, A., Confidence limits for an unknown distribution function, Annals of Mathematical Statistics, 1941, 12, 461-463.

*T 23. Kolmogoroff, A., Sulla determinazione empirica di una legge di distribuzione. Giornale dell' Instituto Italiano degli Attuari, 1933, 4, 83-91.

24. Lehmann, E. L., Consistency and unbiasedness of certain nonparametric tests. Annals of Mathematical Statistics, 1951, 22, 165-179.

25. Lehmann, E. L., The power of rank tests. Annals of Mathematical Statistics, 1953, 24, 23-43.

26. Malmquist, S., On certain confidence contours for distribution functions. Annals of Mathematical Statistics, 1954, 25, 523-533.

* 27. Marshall, A. W., A large-sample test of the hypothesis that one of two random variables is stochastically larger than the other. Journal of the American Statistical Association, 1951, 46, 366-374.

T 28. Marshall, A. W., The small sample distribution of $n\omega_n^2$ Annals of Mathematical Statistics, 1958, 29, 307-309.

*TT 29. MASSEY, F. J., A note on the estimation of a distribution function by confidence limits. Annals of Mathematical Statistics, 1950, 21, 116-119.

30. Massey, F. J., A note on the power of a non-parametric test. Annals of Mathematical Statistics, 1950, 21, 440-443.

31. Massey, F. J., Correction to "A note on the power of a nonparametric test." Annals of Mathematical Statistics, 1952, 23, 637-638.

T 32. Massey, F. J., Distribution table for the deviation between two sample cumulatives. Annals of Mathematical Statistics, 1952, 23, 435-441.

*T 33. MASSEY, F. J., The distribution of the maximum deviation between two sample cumulative step functions. Annals of Mathematical Statistics, 1951, 22, 125-128.

T 34. Massey, F. J., The Kolmogorov-Smirnov test for goodness of fit. Journal of the Americal Statistical Association, 1951, 46, 68-78.

TT 35. Miller, L. H., Table of percentage points of Kolmogorov statistics. Journal of the American Statistical Association, 1956, 51, 111-121.

36. Rosenblatt, M., Remarks on a multivariate transformation. Annals of Mathematical Statistics, 1952, 23, 470-472.

37. Simpson, P. B., Note on the estimation of a bivariate distribution function. Annals of Mathematical Statistics, 1951, 22, 476-477.

38. Smirnov, N. V., Sur la distribution de ω^2 (criterium de M. R. v Mises). Comptes Rendus (Paris), 1936, 202, 449-452.

T 39. Smirnov, N. V., Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, 1948, 19, 279-281.

**TT 40. TSAO, C. K., An extension of Massey's distribution of the maximum deviation between two-sample cumulative step functions. Annals of Mathematical Statistics, 1954, 25, 587-592.

41. van der Waerden, B. L., Order tests for the two-sample problem II and III. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A). 1953, 56, 303-310, 311-316.

42. van der Waerden, B. L., Testing a distribution function. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1953, 56, 201-207.

* 43. Wald, A. and Wolfowitz, J., Confidence limits for continuous distribution functions. _Annals of Mathematical Statistics_, 1939, 10, 105-118.

44. Wald, A. and Wolfowitz, J., Note on confidence limits for continuous distribution functions. _Annals of Mathematical Statistics,_ 1941, 12, 118-119.

CHAPTER XII

MULTI-SAMPLE TESTS

Distribution-free tests to detect a differential effect among three or more treatments are often simply generalizations of an analogous distribution-free test for the two-treatment case. In the following chapter, the rank tests for unmatched and matched data are generalizations of the Wilcoxon and sign tests respectively, while the median test generalizes the two sample test of the same name. Most of the remaining tests are at least analogous to, if not direct generalizations from, a two-sample distribution-free test. However, in no case is the parallelism complete. The test statistic may be based on essentially the same sample information, but may take a different form; or its exact probabilities may have been tabled only for the tiniest of sample sizes, asymptotic, approximate formulae being employed to calculate probabilities in all other cases. For these and other reasons, the multi-sample tests appear to have more in common with each other than with the two-sample test which they "generalize." They are therefore presented together in a single chapter.

1. Rank Tests for Unmatched Data

a. Rationale. Suppose that observations have been taken under a variety of conditions, that the observations are continuously distributed but unmatched, and that it is desired to test whether or not the observations recorded under the various conditions all belong to the same population. Rank the observations from 1 to N, where N is total number of observations recorded under all conditions. Now construct a table with C columns, representing conditions, and with a number of rows equal to the greatest number of observations recorded under a single condition. Enter each rank under the appropriate column paying no attention to the row into which it happens to fall. Let n_i represent the number of entries, i.e., the number of occupied cells, in the i^{th} column, and let R_i represent the sum of the ranks in the i^{th} column. The average rank entry in the entire table is $(N+1)/2$, and, if the null hypothesis is true, it is also the "population" average for the n_i rank entries in the i^{th} column. The expected column sum for the i^{th} column is therefore $n_i(N+1)/2$. Let S represent the sum of the squared deviations of the column sums from their expected values, then $S = \sum_i \left[R_i - \frac{n_i(N+1)}{2} \right]^2$.

For a given table, N cells are occupied. There are N ! ways in which the ranks from 1 to N could have been assigned to these N cells by chance, and if chance is the only determining factor each of these ways is equally likely. Therefore, to determine the probability of an S as great or greater than that obtained, one need only find the number of the N! tables which yield such an S and divide by N! This method however involves excessive computation. The n_i observations in the i^{th} column can be permuted in $n_i!$ ways without changing R_i and therefore without affecting the value of S. For each such permutation, the within-column entries of every other column can be likewise permuted, so there are $\prod_{i=1}^{C} (n_i!)$ ways of permuting within-column entries without affecting the value of S. Therefore, one may save labor by confining his attention to the $\frac{N!}{\prod_{i=1}^{C}(n_i!)}$ tables which can be formed by permuting entries from one column to another. If there are the same number of entries in every column, permutations which merely interchange entire columns of entries do not change S, and since there are C! such column permutations possible for each "table" (i.e., for any given permutation) further labor can be saved by taking as one's population of tables only the $\frac{N!}{C!\prod_{i=1}^{C}(n_i!)}$ tables whose permutations exclude permutations of entire columns and permutations of entries within columns. In either case the cumulative probability of a given value of S is simply the proportion of the restricted population of "tables" which yield an S equal to or greater than the given value. Exact probabilities have been calculated for S and for a statistic, H, which equals $\frac{12}{N(N+1)}\sum_{i=1}^{C}\frac{[R_i - \frac{n_i(N+1)}{2}]^2}{n_i}$ and which is equivalent to S, since the n_i are parameters for the exact tables of both S and H.

The calculations required for the exact method become unwieldy and impractical at very modest sample sizes, at which point approximations must be relied upon. Owing to the effect described in the Central Limit Theorem, a column mean or sum tends to become normally distributed as the number of observations, n_i, upon which it is based increases (assuming C fixed) or, perhaps to a somewhat lesser degree, as the number, C, of different values an observation can assume increases (assuming n_i fixed). Therefore, roughly speaking, the tendency to normality generally increases with increasing N. If the null hypothesis is true, each column mean comes from a population of "column means" whose mean is $\frac{N+1}{2}$ and whose variance is $\frac{N^2-1}{12n_i}\left(\frac{N-n_i}{N-1}\right)$.

(The variance of a distribution of means is $\frac{\sigma^2}{n}(\frac{N-n}{N-1})$, where σ^2 is the population variance, which in the case of sampling without replacement from the population of integers from 1 to N is $\frac{N^2-1}{12}$, and where N and n are the respective sizes of the population and of the sample.) Therefore, if the null hypothesis is true and if N is large enough for the i^{th} column mean to have an essentially normal distribution,

$$\frac{\frac{R_i}{n_i} - \frac{N+1}{2}}{\sqrt{\frac{N^2-1}{12n_i}\left(\frac{N-n_i}{N-1}\right)}}$$

is a standardized normal deviate with zero mean and unit variance. The sum of C squared standardized normal deviates has a chi square distribution with C degrees of freedom <u>if the deviates are independent.</u> In the present case, of course, they are not independent: if C-1 of the R_i are known, the remaining one can be obtained by subtraction from N. However, by making mathematical allowance for the correlation, the above approach can, with a slight modification, be used to obtain a test statistic,

$$H= \frac{N-1}{N}\sum_{i=1}^{C}\frac{\left(\frac{R_i}{n_i}-\frac{N+1}{2}\right)^2}{\frac{N^2-1}{12n_i}}$$

which is distributed approximately as chi square with C-1 degrees of freedom. An equivalent formula, which is more efficient for computation, is

$$H = -3\,(N+1) + \frac{12}{N(N+1)}\sum_{i=1}^{C}\frac{R_i^2}{n_i}.$$

b. <u>Null Hypothesis</u>. The a priori probability that a given rank will belong to an observation in the i^{th} column is n_i/N. This will be the case if all N observations are members of the same population, i.e., if conditions have not affected observations differentially, and if all assumptions are met.

c. Assumptions. Observations have been drawn randomly and independently from continuously distributed populations (which would be identical if conditions had equal effects.) If the approximate test is used it must be further assumed that no n_i is very small, i.e., that in every column there are enough observations so that, in accordance with the Central Limit Theorem, the mean of the n_i ranks will be essentially normally distributed about the grand mean of $\frac{N+1}{2}$.

d. Treatment of Ties. When all the observations forming a tied group lie in one column of the table, ties may be resolved arbitrarily. In all other cases a conservative test calls for ties to be resolved in the manner least conducive to rejection of the null hypothesis; however, probability error will be minimized in the long run if, instead, such ties are assigned the midrank of the tied-for ranks. If the latter method is employed and if the large sample (i.e. approximate) version of the test is used, the mathematical effect of ties can be compensated for by calculating H from the following formula:

$$H = \frac{-3(N+1) + \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i}}{1 - \frac{\sum T}{N^3 - N}}$$ where $T = t^3 - t$ and t is the number of observations tied for the same rank.

e. Efficiency. When both tests are applied to populations having normal distributions differing only in location (and therefore having equal variances) the H test has, relative to the F test of analysis of variance, an asymptotic relative efficiency of $3/\pi$ or .955. Under the same circumstances, it is more efficient than the Brown-Mood median test which has an A.R.E. of 2/3 relative to the H test. If, in the above, "uniform distribution" is substituted for "normal distribution", the A.R.E. of the H test relative to the F test becomes 1.00 and that of the median to the H test becomes 1/3. The

A.R.E. of the H test relative to the F test can exceed 1.00 for certain types of distribution (2). However, if the distributions have identical shapes, it cannot fall far below 1.00. The finding of Hodges and Lehmann concerning the efficiency of the Wilcoxon test relative to the t test, applies also to the efficiency of the H test relative to the F test: If samples are from continuous distributions, differing only in location, the A.R.E. of the H test relative to the F test can never fall below .864.

The H test is consistent against translation alternatives (2). More generally, it is consistent if for some one of the C populations the probability that a randomly selected observation from that population exceeds a randomly selected observation from among all C populations is some value other than 1/2 (19, 20). For example, the test is not consistent if the C populations are symmetrical with equal means but unequal variances, i.e., rejection of the hypothesis of identical populations cannot be assured by taking infinite sized samples.

f. Application. Suppose that speed of reading is to be tested under three degrees of illumination. Nine subjects are selected at random from a common population, and three subjects are randomly assigned to each condition of illumination. Due to some misadventure one subject fails to complete the experiment. Let the data be as shown below, the first table giving the raw scores and the second one showing their ranks.

	Condition	
A	B	C
22	36	39
31	37	44
35		51

	Condition	
A	B	C
1	4	6
2	5	7
3		8

Calculating $H = -3\,(N+1) + \dfrac{12}{N\,(N+1)} \displaystyle\sum_{i=1}^{C} \dfrac{R_i^{\,2}}{n_i}$, we obtain

$H = -3\,(8+1) + \frac{12}{8(8+1)}\left[\frac{6^2}{3} + \frac{9^2}{2} + \frac{21^2}{3}\right] = 6.25$, which is found, by consulting Kruskal and Wallis' tables, to have a probability of .011. This probability could easily have been obtained without the use of tables: There are 8 ranks and 8! ways in which they could have been assigned to the 8 cells of the above table. However permutations among the 3 cell entries in column A, or the 2 in column B, or the 3 in column C are of no interest nor are permutations of the entire set of entries under column A with those under column C. Therefore, only $\frac{8!}{3!\ 2!\ 3!\ 2!}$ or 280 tables need concern us. Of these 280 tables, only 3 yield values of $\sum \frac{R_i^2}{n_i}$ as great or greater than the value actually obtained. They are as follows:

Condition			Condition			Condition		
A	B	C	A	B	C	A	B	C
3	1	6	1	4	6	1	7	4
4	2	7	2	5	7	2	8	5
5		8	3		8	3		6

The probability of a table as extreme or more extreme than that obtained is therefore 3/280 or .011.

g. Discussion. Wilcoxon's two-sample test for unmatched data assigned ranks to observations, irrespective of the sample to which they belonged, then applied Fisher's method of randomization to the rank sums of the samples. White extended the test to samples of unequal size. The present test is a generalization of the Wilcoxon-White test to the multi-sample case, the procedure differing mainly in that the sum of the squared deviations of rank sums from their expected values, (or the equivalent) has, in effect, replaced the rank sum as the test statistic. The Wilcoxon form of the test, requiring equal-sized samples, has been generalized by Rijkoort (29) whose test statistic is the value S, defined under "Rationale". Kruskal and Wallis (19, 20) have generalized the White form of the test which permits samples of unequal size; their test statistic is H.

The Mann-Whitney form of the "Wilcoxon test" applies to unequal as well as equal-sized samples and does not use rank sums as the test statistic. Instead it employs the statistic, U, which is the number of times a Y-sample observation precedes an X-sample observation when observations from the two samples are arranged in a single sequence in order of increasing size. The Mann-Whitney test is a special case of the form of Kendall's rank correlation test (tabled by Sillito) which takes exact account of ties in one ranking. Certain multisample generalizations of the Mann-Whitney test are mathematically equivalent to this form of Kendall's test. Observations may be regarded as having two characteristics: their value and the sample to which they belong. All observations are ranked as to value, and this is the untied ranking. If ranked according to the other characteristic, the result is a ranking containing ties, all of the observations in a given sample being tied for that sample's rank. The rank correlation test then tests whether or not the tied and untied rankings are correlated, i.e., whether or not value ranks are systematically related to sample-category ranks (roughly, it tests whether or not observation values are systematically related to their sample categories).

In generalizations of the Wilcoxon and White forms of the "Wilcoxon" test, the alternative to the null hypothesis is simply that samples differ. Specifically the alternative hypothesis is that the average rank of observations in one or more unspecified columns differs in a real nonchance way from the average rank for the entire table. In generalizations of the Mann-Whitney (and, therefore, "modified Kendall") form of the test, however, the alternative hypothesis is much more specific. It states that observation-value ranks are correlated with their sample ranks and therefore specifies the order of arrangement of the samples. Thus, speaking roughly, it states that observations of intermediate size tend to lie in those samples "assigned" intermediate ranks (and therefore represented by the middle columns) and that either small observations tend to lie in samples with low rank (the left hand columns) and large observations in samples with high rank, or the reverse if "negative" correlation is suspected. The multisample generalizations of the Wilcoxon-White and Mann-Whitney forms are therefore quite different in their applications. The probability tables for the two forms are constructed for different rejection regions, the rejection region for the Mann-Whitney form being taken so as to maximize the probability of rejection when a specific alternative hypothesis is true. Furthermore the test statistic for generalizations of the Mann-Whitney test is based upon inversions rather than rank sums and can assume a greater number of gradations of value than can generalizations of the Wilcoxon or White tests.

Multisample generalizations of the Mann-Whitney test have been proposed, and their exact small sample probabilities tabled, by Terpstra (43) and by Whitney (52). Multisample tests equivalent to Kendall's rank correlation S, with exact allowance for ties in one ranking, have been developed by Krishna-Iyer (18), Terpstra (45) and Jonckheere (12), exact small sample probabilities having been tabled by the last two authors. These tests, in effect, require that "columns" be arranged in an order implied by the alternative hypothesis; they then test whether or not this order bears a "chance" relationship to the rank order of the observations in the table so constructed.

h. Tables. Exact probabilities for H have been tabled (20, 21, I-43) for the case of three samples, none of which contains more than five observations, i.e., $C=3; n_1, n_2, n_3 \leqq 5$. (Samples not necessarily of equal size). Exact probabilities for S have been tabled (29, 21) for the cases in which the number of samples is 3, 4 or 5, each sample containing an equal number of observations (2, 3, 4 or 5 in the first case, 2 or 3 in the second, and 2 observations in the third case in which there are 5 samples).

Various approximations exist for cases not covered by the exact tables. As indicated under Rationale, H is distributed approximately as χ^2 with C-1 degrees of freedom, and so is

$\frac{12(C-1)S}{(N+1)(N^2 - \sum_i n_i^2)}$. The chi square approximation is the easiest to use, but it is not the best. Closer approximations are discussed in (19, 20, 29). A nomogram for obtaining probability levels for the H or S tests is given in (30).

i. Sources. 2, 12, 18, 19, 20, 21, 29, 30, 43, 44, 45, 50, 52, I-43.

2. Rank Tests for Matched Data

a. Rationale. Suppose that each of m subjects has performed under each of n conditions and that one desires to test whether or not the various conditions have equal influence upon performance. Let an m x n table be constructed with n columns, representing conditions, and m rows, representing subjects. Rank each subject's performance under the n conditions, assigning a rank of 1 to the smallest score, 2 to the next smallest, etc., and n to the largest. Then record each rank in the appropriate cell of the m x n table. The cell entries in each row of the table constitute one permutation of the sequence of integers from 1 to n. There are n! possible permutations for a given row. For each such permutation, there are n! ways of permuting a second specified row, etc. Since there are m rows, there are $(n!)^m$ different "tables" which can be obtained by permuting cell entries within rows.

Now sum the cell entries in each column. The average cell entry in a row is $\frac{n+1}{2}$, so the average column sum is $m(\frac{n+1}{2})$. From each column sum subtract $m(\frac{n+1}{2})$ to obtain the deviation of the column sum from the value expected if conditions have equal effects upon performance. Square each deviation and sum the squared deviations. Call this sum S.

For each of the $(n!)^m$ possible tables there will be a corresponding value of S (some tables, of course, yielding the same S). Therefore the exact probability for an S equal to or greater than that obtained is simply the number of the $(n!)^m$

different possible tables which yield such values of S, divided by $(n!)^m$. (Some of the $(n!)^m$ possible tables differ only in that entire columns are interchanged. Since there are n columns, there are n! variations of any given table which can be effected simply by transposing columns. All of these n! variations, of course, yield the same S. Therefore computations can be considerably lessened by counting critical values of S only from the $(n!)^{m-1}$ tables none of which can be obtained by transposing columns of another table in the set. If an S as great or greater than that actually obtained could have been obtained from N of the $(n!)^m$ "unrestricted" tables and from N' of the $(n!)^{m-1}$ "restricted" ones, then $N/(n!)^m$ and $N'/(n!)^{m-1}$ are equal, and both give the exact probability sought.)

When m and n are small exact probabilities can be calculated in the manner indicated above. Such exact probabilities have been tabled for S and for a statistic, $\chi^2 r$, which equals $\frac{12S}{mn(n+1)}$ and is therefore equivalent to S when exact tables are used.

Owing to the effect described in the Central Limit Theorem, the distribution of column means approaches a normal distribution as the number of rows increases, thus making possible an approximate test when m and n exceed the values given for them in the exact tables. The mean and variance of a single table entry are $\frac{n+1}{2}$ and $\frac{n^2-1}{12}$ respectively. The mean of the entries is also the mean of the column means. The variance of a mean of m observations is 1/m times the variance of the individual observations upon which the mean is based. Therefore the variance of a column mean is $\frac{n^2-1}{12m}$. If $\overline{R}_j$ is the mean of the ranks in the jth column, then $\frac{\overline{R}_j - \frac{n+1}{2}}{\sqrt{\frac{n^2-1}{12m}}}$ is, for large values of m, approximately a standardized normal deviate with zero mean and unit variance. The sum of the squares of n independent standardized normal deviates is distributed as chi square with n degrees of freedom. However, the n column means are not independent; knowing n-1 of them, the remaining mean can be obtained by subtracting their sum from $\frac{n(n+1)}{2}$. If one mean is "ignored", however, the remaining n-1 means may be regarded as practically independent. Therefore, n-1 of the column means could be selected at random and used to calculate n-1 standardized normal deviates the sum of whose squared values would be distributed as chi square with n-1 degrees of freedom. This approach, however, is objectionable because the "information" contained in the arbitrarily discarded mean is ignored. The solution favored by Friedman (11) is to find the sum of the squares of all n standardized normal deviates, divide by n to obtain the average squared standardized normal deviate , then multiply this by n-1 to obtain a simulated sum of n-1 squared standardized normal deviates which, nevertheless, takes all n of the deviates into account. The resulting value, $\frac{n-1}{n}\sum_1^n \frac{(\bar{R}_j - \frac{n+1}{2})^2}{\frac{n^2-1}{12m}}$, is distributed approximately as chi square with n-1 degrees of freedom. For computational purposes, it is easier to use the equivalent formula

$$X_r^2 = -3m(n+1) + \frac{12}{mn(n+1)} \sum_1^n R_j^2 \text{ with n-1 degrees of freedom,}$$

where R_j is the sum of the ranks in the j^th column and X_r^2 symbolizes a modified X^2 which has approximately the chi square distribution.

b. Null Hypothesis. For each row, each of the n! permutations of the ranks 1 to n was equally likely to be the sequence of cell entries recorded. This will be the case if conditions have equal influence upon scores (so that variations in a subject's performance are random) and if all assumptions are true. Note: the null hypothesis does not imply that the observations in different rows come from the same population.

c. Assumptions. Observations upon a subject are randomly selected (usually it is also assumed that subjects are randomly selected), rows are independent, i.e., one subject's performances are uninfluenced by the performance of any other subject, and within a single row there are no tied ranks (thus, if "performance" is not intrinsically in rank form, it is assumed to be continuously distributed). If the approximate test is used, it must be further assumed that m, the number of rows, is large enough so that, in accordance with the Central Limit Theorem, the mean of the m ranks in a column will be essentially normally distributed about the grand mean of $\frac{n+1}{2}$.

d. Treatment of Ties. The conservative method of dealing with within-row ties is to distribute the tied-for ranks to tied cells in such a way as to minimize S. In order to minimize error in the long run, give each of the within-row ties the average of the tied-for ranks.

e. Efficiency. When n=2, the present test is equivalent to the sign test which has an asymptotic efficiency of .637 relative to Student's t test. Therefore, when n=2 and m is infinite, the present test has an efficiency of .637 relative to Student's t (11). This is presumably the lowest efficiency value assumed by the test since the efficiency of the sign test increases with decreasing sample size and since when n=2 the ranks in effect designate only "smaller" versus "larger", while, with increasing n, finer and finer gradations of discrimination are possible. Thus, with increasing n, ranks simulate more and more closely the gradations of measurement characteristic of continuously distributed original scores and efficiency should approach that of tests based on such scores. At the other extreme, when m=2, the test is equivalent to the rank difference correlation test shown by Hotelling and Pabst to have an asymptotic efficiency of .912 relative to the parametric test for correlation. This then is the efficiency of the present test when m=2 and n is infinite (11). It seems reasonable to conclude, therefore, that the present test when applied to normally distributed original scores has, relative to parametric tests, an efficiency no smaller than .637 and generally considerably higher. (If either n=2 and m is very small, or if m and n are both quite large, one would expect an efficiency close to 1.00).

f. Application. Each of three subjects performs a well learned task three times, each time under the influence of a different drug. Performance is timed and the experimenter wishes to test the hypothesis that no subject's performance times were influenced more by one drug than by another.

TIME SCORES

	Drug		
Subject	I	II	III
A	4.76	1.30	7.91
B	14.51	10.27	35.84
C	82.11	82.09	82.14

TIME-SCORE RANKS

	Drug		
Subject	I	II	III
A	2	1	3
B	2	1	3
C	2	1	3
SUM	6	3	9

The original scores are shown above, a second table substituting ranks for scores. For the latter table, the average column sum is $m\left(\frac{n+1}{2}\right)$ or 6, so the deviations of the column sums from their mean value are 0, -3, and 3. Squared these become 0, 9, and 9 and their sum S is 18. Consulting Kendall's exact tables it is found that an S of 18 has a chance probability of .028 of being equalled or exceeded. (The test is one-tailed since S can only be positive and since very small values of S only indicate unlikely degrees of "agreement" with the null hypothesis.)

The same result could have been calculated without resort to tables. There are 3! ways of assigning the integers 1, 2, and 3 to the three cells in row B and for each of these permutations, there are 3! ways of permuting the ranks in row C. Thus there are 3! x 3! = 36 tables which can be constructed without altering the rankings in row A. For each of these tables there is a set of column sums and a corresponding value of S. The actually obtained set of column sums, however, differ maximally and can be obtained in only one way if A's ranking is held constant (any permutation of B's or C's ranks bring the sums closer together and reduce S). Therefore the obtained table yields the maximum value of S which can be obtained in only one of 36 tables, and the probability of an S equal to or greater than that obtained is 1/36 or .028. The hypothesis of equal drug effects can therefore be rejected at beyond the .05 level of significance.

g. Discussion. Pitman (27) extended Fisher's Method of Randomization for matched observations from the two treatment case to the case of multiple treatments. The present test, when used as an exact test, differs from that proposed by Pitman only in that ranks have been substituted for original observations and the test statistic is S rather than the F ratio. The further extension of the test from its present requirement of one observation per cell to the case where any cell can be empty or contain any positive number of observations has been discussed by Benard and van Elteren (3).

The present test is exact only when probabilities are obtained by the Method of Randomization. When they are obtained from the Z or X^2 distribution, (see "Tables") they are approximate. For values of m and n slightly larger than those for which exact probabilities of S have been tabled, the approximate probabilities obtained by using the Z tables are reasonably close to the true values at the .05 or .01 levels of significance; however, the .001 level of significance should be avoided. The tails of the distribution of S are very irregular when m and n are in this region.

It is to be noted that the test does not assume homogeneity of rows. The "subjects" may belong to different populations and their absolute performance scores under a given condition may differ tremendously. Furthermore, the variability of performance under the various conditions may differ vastly from one subject to another. The test is not designed to detect such effects. It essays merely to detect any systematic tendency for performance under one condition to be superior to that under another condition. It will fail if such a tendency exists in some rows but is balanced by an opposite tendency in other rows. Therefore, while homogeneity of rows is not assumed by the test, it will generally be desirable to select subjects from the same population. If this is not done, and if such subjects are not selected randomly so as to be "representative" of their populations, the results of the test will, in a sense, be peculiar to the group actually tested.

The preceding test can be used to test for interactions (53). The method can be best explained in terms of an example. Suppose that four subjects have performed under each of three conditions, I, II and III, of one variable and have done so under each of two conditions, A and B, of another variable. (The significance of each variable alone can be tested, by collapsing data over the other variable and performing the test as described earlier.) It is desired to test whether the two variables interact. Let the data be as shown below:

BLOCK	ROW (Subject)	COLUMN I	II	III
A	1	15.4	26.9	27.8
	2	14.6	25.9	28.7
	3	8.3	14.2	12.0
	4	5.9	19.9	20.3
B	1	9.2	15.1	18.7
	2	5.1	10.2	15.4
	3	4.9	8.2	6.1
	4	11.5	12.5	29.1

Now subtract each score in block B from the corresponding score in block A and form the table shown below:

"A" Observations Minus "B" Observations

	COLUMN		
ROW	I	II	III
1	6.2	11.8	9.1
2	9.5	15.7	13.3
3	3.4	6.0	5.9
4	-5.6	7.4	-8.8

The preceding test is then applied to this table in the usual manner as shown below, ranks being substituted for difference-scores.

	COLUMN		
ROW	I	II	III
1	1	3	2
2	1	3	2
3	1	3	2
4	2	3	1

The column sums are 5, 12, and 7, and since the average sum is 8, the squared deviations from the mean sum are 9, 16, and 1 yielding an S of 26 which is significant at the .042 level

If there had been three blocks, A, B and C, two tables would have been constructed, one for the A-B differences and one for the differences, $\frac{A + B}{2}$ - C, or the ultimately equivalent differences, (A + B) - 2C. The statistic $X^2_r = \frac{12S}{mn(n+1)}$ is then calculated for each table after substituting ranks for difference-scores. The sum of these two X^2_r's is distributed approximately as chi-square with 2(n-1) degrees of freedom. With four blocks, A, B, C and D, three tables would be constructed, one each for the differences A-B, A + B-2C, and A+B+C-3D. Ranks would be substituted for difference scores and X^2_r would be calculated for each table. Since each X^2_r is approximately distributed as chi-square with n-1 degrees of freedom, by the additive property of chi-square their sum is distributed as chi-square with 3(n-1) degrees of freedom.

The hypothesis tested is that, except for chance fluctuations, each score in the i^{th} row of one block differs by a constant amount from the corresponding score in the i^{th} row of another specified block. If this is not the case then the influence of columns upon entries of the i^{th} row depends upon blocks and a column-block interaction exists.

Since, in each row, ranks are substituted for original observations, the method is particularly suitable when original data are in intrinsic rank form, each row containing the ranks from 1 to n. This is, in fact, the case when each of m judges ranks each of n things, tied ranks being disallowed. The present test will test the judges' accuracy. Their reliability, i.e., agreement with one another rather than with the "true" ranking, however, is also of some interest and can be tested by means of distribution-free tests originated by Kendall (see "Miscellaneous Distribution-Free Tests") and others (6).

h. Tables. The exact probability that S will equal or exceed a given value has been tabled (14, 15, 16) for the cases: n=3, $2 \leq m \leq 10$; n = 4, $2 \leq m \leq 6$; and n = 5, m=3. Analogous exact probabilities for X^2_r have been tabled (11, I-43) for the cases: n=3,

$2 \leqq m \leqq 9$; $n=4$, $2 \leqq m \leqq 4$; (in the actual notation used P is substituted for n above and n for m above).

For cases not covered by the above tables, close approximate probabilities can be obtained by entering Fisher's Z tables (given in 14) with degrees of freedom $V_1 = n-1-\frac{2}{m}$ and $V_2 = (m-1)V_1$ and with $Z = \frac{1}{2} \log_e \frac{12S\,(m-1)}{m^2(n^3-n)-12S}$ or, somewhat more accurately, with Z corrected for continuity, $Z = \frac{1}{2} \log_e \frac{12(S-1)(m-1)}{m^2(n^3-n)-12(S-3)}$.

Using the above formula, with correction for continuity, tables have been prepared which give values of S significant at the .05 and .01 levels for the cases $3 \leqq n \leqq 7$, m = 3, 4, 5, 6, 8, 10, 15 or 20 (10,14, I-43). Using the identity $X^2_r = \frac{12S}{mn(n+1)}$ these tables can be "translated" into analogous tables for X^2_r. This,in effect, has been done (10), the tables being expanded to cover the additional cases m = 100 and m = infinity. A nomogram based upon still another approximation is available in (30).

If the statistic X^2_r is used instead of S, close approximate probabilities can be obtained by substituting $mn(n+1)X^2_r/12$ for S in one of the formulae, given above, for Z, and then consulting the Z tables. A less close approximation to exact probabilities can be somewhat more readily obtained by entering the chi-square tables with n-1 degrees of freedom and with $X^2 = \frac{12S}{mn(n+1)}$, or, corrected for continuity $X^2 = \frac{12m(n-1)(S-1)}{m^2(n^3-n)+24}$.

When Z or X^2 tables are used to obtain probabilities, corrections for ties may be made. These are given by Kendall (14).

The simplest procedure appears to be to correct X_r^2 for ties and then find the S corresponding to this value of X_r^2. Corrected for ties, $X_r^2 = \dfrac{S}{\dfrac{mn(n+1)}{12} - \dfrac{\sum T}{n-1}}$ where $T = \frac{1}{12}\sum (t^3-t)$, t being the number of observations in a particular row which are tied for the same rank, the summation of (t^3-t) occurring over all tied-for ranks in that particular row and the summation of T occurring over all rows.

i. Sources. 3, 6, 9, 10, 11, 13, 14, 15, 16, 27, 30, 32, 34, 35, 40, 41, 42, 51, 53, 54.

3. Median Tests

a Rationale. Suppose that (continuously distributed) observations have been taken under C experimental conditions and that it is desired to test whether or not the conditions have equal effects. Let n be the total number of observations and a be the number of those observations which lie above the grand median, M, and let n_i be the number of observations taken under the ith experimental condition and a_i be the number of those n_i observations which lie above the grand median for all observations.

a_1	a_2	-----	a_{c-1}	a_c	a
n_1-a_1	n_2-a_2	-----	$n_{c-1}-a_{c-1}$	n_c-a_c	n-a
n_1	n_2	-----	n_{c-1}	n_c	n

Suppose that the conditions do have equal effects so that all n observations are from a common, continuously distributed, population. If P is the proportion of the population lying above the grand median of the n observations, the a priori probability that a_i of the n_i observations taken under the i^{th} condition will exceed M is $\binom{n_i}{a_i} P^{a_i} (1-P)^{n_i - a_i}$, and the a priori probability that under successive conditions the number of observations above the median will be $a_1, a_2, \ldots, a_c$ is the product $\prod_{i=1}^{c} \binom{n_i}{a_i} P^{a_i} (1-P)^{n_i - a_i}$. However, the value, M, used to dichotomize the data into the frequency categories a_i and $n_i - a_i$, is a sample, not a population, median, i.e., it was determined a posteriori. Therefore, the probability we seek is the conditional probability of cell entries $a_1, a_2, \ldots, a_c$, given that their marginal total is a, and this is obtained by dividing the a priori probability $\prod_{i=1}^{c} \binom{n_i}{a_i} P^{a_i} (1-P)^{n_i - a_i}$ by the a priori probability of the marginal totals, which is $\binom{n}{a} P^{a} (1-P)^{n-a}$. In the resulting fraction the terms containing P cancel out leaving

$$\frac{\prod_{i=1}^{c} \binom{n_i}{a_i}}{\binom{n}{a}}$$

as the point probability for the obtained table. The significance level for a given table is obtained by cumulating the point probabilities for all tables as extreme or more so. However, with increasing values of n_i or C calculations are likely to become prohibitively laborious. Fortunately, when $n \geqq 20$ and all $n_i \geqq 5$ a fairly good approximate test can be performed by calculating

$$\frac{n\,(n-1)}{a\,(n-a)} \sum_{i=1}^{c} \frac{\left(a_i - \frac{n_i a}{n}\right)^2}{n_i}$$

which is distributed very nearly as chi-square with C-1 degrees of freedom.

b. Null Hypothesis. The probability that an observation will be above the grand sample median, M, is independent of the experimental condition under which the observation was taken. This will be the case if conditions have equal effects and if all assumptions are met.

c. Assumptions. Sampling is random, observations are independent, and there are no tied observations, i.e., the sampled populations are continuously distributed. If the large sample approximation is used, then all of the assumptions of chi-square are also introduced.

d. Treatment of Ties. Tied observations are no problem unless they are tied with the median. In this case, if the proportion of such ties is small, the following procedure is recommended. Either (a) resolve all ties in the manner least conducive to rejection of the null hypothesis, or (b) under each condition separately count half of the observations tied with the grand median as above it, half as below it, and treat an odd tie as outlined in (a) above.

e. Efficiency. When both tests are applied to populations having normal distributions, differing only in location (and therefore having equal variances) the median test has, relative to the F test of analysis of variance an asymptotic relative efficiency of $2/\pi$ or .637. Under the same circumstances it has an A.R.E. of 2/3 relative to Kruskal and Wallis' H test. If, in the above, "uniform distribution" is substituted for "normal distribution", the median test has A.R.E. of 1/3 relative to the F test and also relative to the H test. For other types of distribution, the A.R.E. of the median test relative to either the F or H tests can be less than, equal to, or greater than, 1, depending upon the particular distribution to which applied (2).

The median test is consistent against translation alternatives (2).

f. Application. Suppose that 16 rats have been randomly selected from a common population and randomly divided into three groups. Each group is administered a different drug after which the time to run a maze is measured for each rat. The null hypothesis is that the three drugs have equal effects upon maze running ability. Let the data be as shown below.

Maze Running Times Under

Drug A	Drug B	Drug C
267	269	215
271	283	231
285	288	233
299	302	252
304	306	255
		264

The grand sample median lies between 269 and 271, therefore, in terms of frequencies of observations above and below the median the above table becomes:

	A	B	C	Totals
Above Median	4	4	0	8
Not Above Median	1	1	6	8
	5	5	6	16

The first row of tables as extreme or more so, and their point probabilities are given below:

A	B	C	Point Probability = $\Pi \binom{n_i}{a_i} / \binom{n}{a}$
5	3	0	10/12870
3	5	0	10/12870
4	4	0	25/12870
1	1	6	25/12870
2	0	6	10/12870
0	2	6	10/12870

The cumulative probability is 90/12870 or .007. Recalculating the probability using the chi-square approximation, we have

$$X^2 = \frac{n(n-1)}{a(n-a)} \sum_{i=1}^{c} \frac{\left(a_i - \frac{n_i a}{n}\right)^2}{n_i} = \frac{16 \times 15}{8 \times 8} \left[\frac{(4-(\frac{5}{16})\, 8)^2}{5} + \frac{(4-(\frac{5}{16})\, 8)^2}{5} + \frac{(0-(\frac{6}{16})\, 8)^2}{6} \right] = 9.$$

Entering the chi-square tables with $C-1=2$ degrees of freedom, a chi-square of 9.00 is found to have a probability just slightly larger than the .01 level of significance. Both methods give a probability in the neighborhood of .01, but it is clear that the approximation is not impressively close to the true value.

g. <u>Discussion.</u> Mood (24) and Brown and Mood (5) have outlined median tests for cases analogous to those encountered in a two way analysis of variance. Exact tests are theoretically possible, for these cases, but actually impractical because of the laborious computations involved. The user is therefore practically forced to ignore the exact probability formulae given by Mood, relying instead upon test statistics which have approximately the chi-square distribution. The tests are described by the authors referenced above and will only be briefly outlined here.

Test for main effects in a two-factor experiment with one observation per cell: To test for column effects find the median for each row and count the number of observations in each column which exceed their respective row medians. Let a_i be the number of such observations in the i^{th} column, let there be r rows and c columns, and let $a = c/2$ if c is even or $\frac{c-1}{2}$ if c is odd (note change in definitions of a_i and a). Then each row contains a observations exceeding the row median, and the table

a_1	a_2	-----	a_{c-1}	a_c	ra
$r-a_1$	$r-a_2$	-----	$r-a_{c-1}$	$r-a_c$	r(c-a)
r	r	-----	r	r	cr

contains ra such observations. If columns have equal effects, the expected number of observations, in each column, which exceed their respective row medians is ra/c and the value

$$\frac{c(c-1)}{ra(c-a)}\sum_{1}^{c}(a_i-\frac{ra}{c})^2$$ is asymptotically distributed as chi-square with

c-1 degrees of freedom. Since the expected frequency, ra/c is simply the number of above-row-median observations in the entire table divided by the number of columns, its use implies that there are no interaction effects and therefore introduces this assumption (unless one or both factors have randomly chosen levels). An additional assumption is that all observations have distributions which are identical except for location. Every observation within a row must, before sampling, have had equal probability, under the null hypothesis, of exceeding the row median. This means that an equal proportion of each observation's population distribution must lie above the row median. Since the row median is not fixed, but is a variable, sample value, the above requirement is certain to be fulfilled only if every observation within the same row has a distribution of the same shape. In testing for row effects, a similar argument requires that observations within the same column have distributions of the same form. In testing for both row and column effects, therefore, since a row observation is also a column observation, all observations must be distributed identically except for location. Naturally the assumptions listed under (c) must also be made.

Tests for various effects in a two-factor experiment with h observations per cell: Again it is assumed that observations have distributions which are identical except for location. A test analogous to the "analysis of variance" test for main effects against interaction can be made by performing the test outlined in the preceding paragraph using cell medians as "observations". Another useful test is the joint test for main effects and interaction. It tests the "hypothesis that a factor has no effect whatever, either in main effects or in interaction effects". Let a_{ij} be the number of observations, in the cell formed by the i^{th} row and the j^{th} column, which exceed the median of the ch observations in the i^{th} row;

and let a = ch/2, if ch is even, or $\frac{ch-1}{2}$ if ch is odd. Then if,

as hypothesized there are no interaction or column effects, the expected number of observations in a single row exceeding a row median is a, and the expected number of observations in a single cell exceeding the corresponding row median is a/c. Thus

$\frac{c\,(ch-1)}{a\,(ch-a)}\sum_{i,j}(a_{ij}-\frac{a}{c})^2$ is distributed approximately as chi-square

with r(c-1) degrees of freedom. Analogous to the test of main effects against deviations the following test can be performed if interactions can be assumed to be zero. Let a_i be the number of the rh observations in the i^{th} column which exceed their row

medians, and let $a = \frac{ch}{2}$, if ch is even, or $\frac{ch-1}{2}$ if ch is odd.

Then there are rh observations in a column, and, $\frac{ra}{c}$ of them would be expected to exceed their respective row medians. Thus

$$\frac{c(ch-1)}{ra(ch-a)}\sum_i (a_i - \frac{ra}{c})^2$$

is distributed approximately as chi-square with c-1 degrees of freedom. Testing for interaction requires that column and row effects be removed by subtraction of column medians from observations followed by subtraction of row medians the process being continued until both columns and rows have zero medians. Let a_{ij} be the number of observations in the ijth cell which exceed its median plus half the number of such observations which equal its median. Let $a_{i.}$ and $a_{.j}$ be a_{ij} summed over columns and rows respectively and let a be the sum over both. Then

$$a^2h\sum_{ij}\frac{(a_{ij} - \frac{a_{i.}a_{.j}}{a})^2}{a_{i.}\ a_{.j}(h-a_{i.}\ a_{.j})}$$

is approximately distributed as chi-square with (c-1)(r-1) degrees of freedom. The test for interactions is "very nearly but not completely distribution-free".

A different approach to median-test analogues of analysis of variance has been taken by Wilson (56), the technique being based upon the fact that a "total" chi-square can be subdivided into component chi-squares with component degrees of freedom (in a sense, the reverse of the additive property of chi-square). In a table with r rows and c columns, let n be the total number of observations, n_{ij} be the number of observations in the cell formed by the i^{th} row and the j^{th} column, $_af_{ij}$ and $_bf_{ij}$ be the number of this cell's observations which are respectively above and below the grand median for the entire table, and n_a and n_b the total number of observations above and below the grand median. Finally let a dot, in place of a subscript, indicate summation over all values of that subscript. Formulae for the total chi-square, X^2_T, the row and column chi-squares, X^2_R and X^2_C are as follows:

$$X^2_T = \sum_i \sum_j \left[\frac{\left({}_af_{ij} - \frac{n_{ij}n_a}{n}\right)^2}{\frac{n_{ij}n_a}{n}} + \frac{\left({}_bf_{ij} - \frac{n_{ij}n_b}{n}\right)^2}{\frac{n_{ij}n_b}{n}} \right] \text{ with rc-1}$$

degrees of freedom, the expected frequencies having been derived from "the null hypothesis that the main effects and interaction effects produce no change in the distribution of scores",

$$X^2_R = \sum_i \left[\frac{\left({}_af_{i.} - \frac{n_{i.}n_a}{n}\right)^2}{\frac{n_{i.}n_a}{n}} + \frac{\left({}_bf_{i.} - \frac{n_{i.}n_b}{n}\right)^2}{\frac{n_{i.}n_b}{n}} \right] \text{ with r-1 degrees}$$

of freedom, $X^2_C = \sum_j \left[\frac{\left({}_af_{.j} - \frac{n_{.j}n_a}{n}\right)^2}{\frac{n_{.j}n_a}{n}} + \frac{\left({}_bf_{.j} - \frac{n_{.j}n_b}{n}\right)^2}{\frac{n_{.j}n_b}{n}} \right]$

with c-1 degrees of freedom, expected frequencies, in both cases, having been obtained from "the null hypothesis that the distributions of scores are identical for all levels of the row or column conditions". Finally an interaction chi-square, X^2_I, is obtained by subtraction, $X^2_I = X^2_T - X^2_R - X^2_C$ with (rc-1) - (r-1) - (c-1) = (r-1)(c-1) degrees of freedom. Computational formulae for extension of the technique to the three-factor case have been published by Alluisi (1).

It has been pointed out that the test compares poorly with the analysis of variance in cases where the assumptions of the latter test have been met (23, 39). Sheffield (33) has objected that an entirely equivalent test can be performed using analysis of variance techniques. Frequencies of "above" or "below median" are treated as scores and their within-cell variance is known to be that of a binomially distributed variate and can therefore be specified a priori. With this information the analysis of variance is conducted upon frequencies. "The implications of these F tests are exactly the same as those of Wilson's X^2 analysis. In fact, since X^2 divided by its df is distributed the same as F for infinite df in the smaller variance, the present F values can be transformed into Wilson's X^2 values by multiplying F by df . . ." Sheffield comments further that, whichever approach is used, severe restrictions are imposed by confining tests to those which can be performed using the single (within-cell) error term which is determinable a priori and therefore distribution-free.

h. Tables. There appear to be no tables for the exact methods, so in these cases probabilities must be computed. Ordinary chi-square tables are used with the approximate methods.

i. Sources. 1, 4, 5, 22, 23, 24, 28, 33, 39, 56.

4. Contingency Tables

Several ingenious distribution-free tests have been devised to examine the significance of effects in an r x c table whose cell entries consist of frequencies rather than "scores" (4, 7, 8, 14, 17, 31, 36, 38, 55).

Suppose that columns are "treatments" whose outcomes are categorized only according to the dichotomy, "success" and "failure", and suppose that rows are "subjects" so that data within a row are matched, each subject receiving all treatments. If a success is obtained on the i^{th} subject for the j^{th} treatment, a 1 is recorded in the ij cell; if the treatment is a failure, a zero is entered in the cell. Let μ_i be the marginal total, i.e., the number of successes, for the i^{th} row, T_j be the marginal total for the j^{th} column, and $\overline{T}$ be the mean column sum. If the number of rows is large, column totals will tend to have normal distributions, and if treatments have equal effects, these distributions will have equal variances and a common mean. As a consequence, the

$$\text{statistic } Q = \frac{c(c-1)\sum(T_j - \overline{T})^2}{c\left(\sum \mu_i\right) - \left(\sum u_i^2\right)}$$ will be distributed approximately as chi-square with c-1 degrees of freedom. This test has been proposed by Cochran (7) as a statistical solution for

the case where matching invalidates the usual chi-square test for contingency tables. This test and certain median tests are special cases of more general tests of dichotomized data outlined by Blomqvist (4).

Suppose that row categories represent gradations or subcategories of a single variable, the level of the gradation or subcategory increasing or decreasing monotonically in progressing from the first to the last row, and suppose that a similar condition exists for columns. A test for association in the contingency table may be regarded then as a test for correlation between the column variable and the row variable. If the 1st column is regarded as having a rank of 1, the 2nd as having a rank of 2, etc., and likewise for rows, then the frequency in the i^{th} row may be considered the number of units tied for a rank of i on the row variable, and similarly the number of units tied in the j^{th} column would be the number tied for a rank of j on the column variable. Finally, the frequency in the i j^{th} cell would be the number of units tied for a rank of i on the row variable which are also tied for a rank of j on the column variable. Thus the situation is analogous to that in which correlation is to be measured between ranked variates when both rankings contain ties. Stuart (38) proposes to calculate Kendall's rank correlation statistic, S, in this case by multiplying each cell frequency (a) positively by the sum of the frequencies in all cells lying below it and to the right, (b) negatively by the sum of the frequencies in all cells lying below it and to the left, (frequencies for cells in the same row, the same column, or above, are ignored): the sum of (a) plus (b), taken over all cells, is S. If the number of rows equals the number of columns, the significance of S can then be tested by techniques taking account of ties in the application of Kendall's test for rank correlation. Otherwise the test can be performed using asymptotic formulae given by Stuart.

5. Tests for a Divergent Population

a. Rationale. Suppose that an experimenter has a sample from each of k continuously distributed populations with identical forms and wishes to test the hypothesis that all populations have the same location against the alternative hypothesis that one population has a larger location parameter than the rest. The sample containing the largest observation is determined, and in it the experimenter counts the number, r, of observations which exceed all observations in all other samples. If n_i is the size of the i^{th} sample and N is the total number of observations in all samples, then there are $n_i(n_i-1) \ldots (n_i-r+1)$ or $n_i!/(n_i-r)!$ ways in which the r largest observations could have been placed in the i^{th} sample and $N(N-1) \ldots (N-r+1)$ or $N!/(N-r)!$ ways in which they could have been located without restriction. The probability that the r largest observations will all be in a preselected sample is therefore

$$\frac{n_i!/(n_i-r)!}{N!/(N-r)!},$$

and the probability that they will all be in some one of the k samples is

$$\Pr(r) = \frac{\sum_{i=1}^{k} n_i!/(n_i-r)!}{N!/(N-r)!}.$$

Since in the derivation it was not required that the (r+1)st largest observation be located in a different sample, the above probability is the probability that r or more of the largest observations will be located in a single sample.

b. Null Hypothesis. The probability that any given one of the r largest observations will be located in a certain sample depends only upon r and the relative size of the sample. This will be the case if all k sampled populations have the same location parameter and if all assumptions are met.

c. Assumptions. Populations are continuously and, except for location, identically distributed. Sampling is random and observations are independent.

d. Treatment of Ties. If the proportion of tied observations is small, ties are a practical problem only if the smallest one of the r largest observations is tied with an observation in a different sample. In this case the simplest solution is to reduce the value of r to the point at which this situation no longer exists. The corresponding probability will be larger than the true probability for the unreduced r, and the test will therefore be conservative.

e. Efficiency. The power of the test has been examined by Mosteller (25) with $r = 3$ for three samples of three observations each from normally distributed and from uniformly distributed populations.

f. Application. In the following table, the four largest observations are all in sample C. Substituting r ▪ 4, $k = 3$, $n_1 = 5$, $n_2 = 5$, $n_3 = 5$ into the

	Sample	
A	B	C
25	27	41
31	35	59
44	39	64
51	48	70
52	57	72

formula given earlier, $\Pr(4) = \dfrac{5!/(5-4)! + 5!/(5-4)! + 5!/(5-4)!}{15!/(15-4)!} =$ 1/91 or .011. This same value could have been obtained by consulting Mosteller's (25) exact tables. The hypothesis of identical populations is therefore rejected. Assuming identical distribution forms, different distribution locations are indicated, and the most reasonable presumption is that the median of population C lies above those of populations A and B.

g. Discussion. Obviously a test which uses as test statistic only the largest observations must be extremely sensitive to both the shape and location of the upper tail of the distribution of the sampled populations. This should be borne in mind when conducting the test. If the assumptions are not fully met, the test may be merely detecting differences in contour-of-upper-tail between distributions with identical locations.

A number of authors (17, 37, 46, 47, 48, 49) have examined tests for divergent populations. Tukey (48, 49) has tabled the probability for the largest column total, i.e. rank sum, when the ranks from 1 to N are randomly distributed among k columns. Both the size and presence of an entry are randomly distributed, i.e., a given column may contain any number of ranks from 0 to N. Tsao (46) has published tables which can be used to obtain the probability for the rank sum of a predesignated column when ranks from 1 to c are substituted for observations matched across rows in a table with c columns and r rows.

h. Tables. Exact tables have been published by Mosteller (25) for the case of equal sized samples ($n_1 = n_2 = n_3 = 3, 5, 7, 10, 15, 20, 25, \infty$) with $2 \leqq k \leqq 6$ and $2 \leqq r \leqq 5$ or 6. Approximate probability tables, appropriate when samples are of unequal size, have been published by Mosteller and Tukey (26). Approximate probability formulae are also given by Mosteller. A simple asymptotic approximation is $Pr\ (r) \approx 1/k^{r-1}$.

i. Sources. 17, 25, 26, 46, 47, 48, 49.

BIBLIOGRAPHY

1. Alluisi, E. A., *Computational formulae for a distribution-free test of analysis-of-variance hypotheses,* Wright Air Development Center Technical Report No. 56-339, July 1956 (Armed Services Technical Information Agency Document No. AD 110445).

2. Andrews, F. C., Asymptotic behavior of some rank tests for analysis of variance. *Annals of Mathematical Statistics,* 1954, 25, 724-736.

* 3. Benard, A. and van Elteren, Ph., A generalization of the method of *m* rankings. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* (A), 1953, 56, 358-369.

* 4. BLOMQVIST, N., Some tests based on dichotomization. *Annals of Mathematical Statistics*, 1951, 22, 362-371.

5. BROWN, G. W. and MOOD, A. M., On median tests for linear hypotheses. *Proceedings Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, 159-166.

T* 6. Cartwright, D. S., A rapid non-parametric estimate of multi-judge reliability. *Psychometrika,* 1956, 21, 17-29.

* 7. Cochran, W. G., The comparison of percentages in matched samples. *Biometrika,* 1950, 37, 256-266.

* 8. van Eeden, Constance and Hemelrijk, J., A test for the equality of probabilities against a class of specified alternative hypotheses, including trend. I & II. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A)*, 1955, 58, 191-198 and 301-308.

9. van Elteren, Ph., The asymptotic distribution for large *m* of Terpstra's statistic for the problem of *m* rankings. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A)*, 1957, 60, 522-534.

T 10. Friedman, M., A comparison of alternative tests of significance for the problem of *m* rankings. *Annals of Mathematical Statistics*, 1940, 11, 86-92.

T* 11. FRIEDMAN, M., The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, 32, 675-701.

* 12. Jonckheere, A. R., A distribution-free k-sample test against ordered alternatives. *Biometrika*, 1954, 41, 133-145.

13. Kendall, M. G., Note on the estimation of a ranking. *Journal of the Royal Statistical Society*, (B), 1942, 105, 119-121.

T 14. KENDALL, M. G., *Rank correlation methods*, 2nd Ed., New York: Hafner, 1955.

T 15. Kendall, M. G., *The advanced theory of statistics* Vol. I., London: Griffin, 1947, 410-421.

T* 16. Kendall, M. G. and Smith, B. B., The problem of *m* rankings. *Annals of Mathematical Statistics*, 1939, 10, 275-287.

T* 17. Kozelka, R. M., Approximate upper percentage points for extreme values in multinomial sampling. *Annals of Mathematical Statistics*, 1956, 27, 507-512.

* 18. Krishna-Iyer, P. V., A non-parametric method of testing K samples. *Nature* (London), 1951, 167, 33.

19. Kruskal, W. H., A nonparametric test for the several sample problem. *Annals of Mathematical Statistics*, 1952, 23, 525-540.

T* 20. KRUSKAL, W. H. and WALLIS, W. A., Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 1952, 47, 583-621.

T 21. Kruskal, W. H. and Wallis, W. A., Use of ranks in one-criterion variance analysis. Vol. 47, No. 267 (December 1952) 583-621, Journal of the American Statistical Association, 1953, 48, 907-911.

* 22. Massey, F. J., A note on a two sample test. Annals of Mathematical Statistics, 1951, 22, 304-306.

23. McNemar, Q., On Wilson's distribution-free test of analysis of variance hypotheses. Psychological Bulletin, 1957, 54, 361-362.

** 24. MOOD, A. M., Introduction to the theory of statistics, New York: McGraw-Hill, 1950, 394-406.

T* 25. MOSTELLER, F., A k-sample slippage test for an extreme population. Annals of Mathematical Statistics, 1948, 19, 58-65.

T 26. Mosteller, F. and Tukey, J. W., Significance levels for a k-sample slippage test. Annals of Mathematical Statistics, 1950, 21, 120-123.

27. Pitman, E. J. G., Significance tests which may be applied to samples from any populations III. The analysis of variance test. Biometrika, 1937, 29, 322-335.

28. Rao, C. R., Advanced statistical methods in biometric research, New York: Wiley, 1952.

T* 29. Rijkoort, P. J., A generalization of Wilcoxon's test. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1952, 55, 394-404.

T 30. Rijkoort, P. J. and Wise, M. E., Simple approximations and nomograms for two ranking tests. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 1953, 56, 294-302.

31. Roy, S. N. and Mitra, S. K., An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. Biometrika, 1956, 43, 361-376.

32. Schultz, F. G., Recent developments in the statistical analysis of ranked data adopted to educational research. *Journal Experimental Education*, 1945, 13, 149-152.

33. Sheffield, F. D., Comment on distribution-free factorial-design analysis. *Psychological Bulletin*, 1957, 54, 426-428.

34. Silvey, S. D., The asymptotic distributions of statistics arising in certain non-parametric tests. *Proceedings Glasgow Mathematical Association*, 1954, 2, 47-51.

35. Stuart, A., An application of the distribution of the ranking concordance coefficient. *Biometrika*, 1951, 38, 33-42.

* 36. Stuart, A., A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 1955, 42, 412-416.

37. Stuart, A., Limit distributions for total rank values. *British Journal of Statistical Psychology*, 1954, 7, 50-51.

* 38. Stuart, A., The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 1953, 40, 105-110.

39. Sutcliffe, J. P., A general method of analysis of frequency data for multiple classification designs. *Psychological Bulletin*, 1957, 54, 134-137.

40. Taylor, F. K., Remarks concerning Willerman's paper on Kendall's W and Sociometric-type ranking. *Psychological Bulletin* 1956, 53, 108.

* 41. Terpstra, T. J., A generalization of Kendall's rank correlation statistic. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* (A)., 1955, 58, 690-696.

42. Terpstra, T. J., A generalization of Kendall's rank correlation statistic II. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* (A). 1956, 59, 59-66.

T* 43. Terpstra, T. J., A non-parametric test for the problem of k samples. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A). 1954, 57, 505-512.

44. Terpstra, T. J., The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A). 1952, 55, 327-333.

T* 45. Terpstra, T. J., The exact probability distribution of the T statistic for testing against trend and its normal approximation. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A). 1953, 56, 433-437.

T* 46. Tsao, C. K., Distribution of the sum in random samples from a discrete population. Annals of Mathematical Statistics, 1956, 27, 703-712.

47. Tsao, C. K., Rank sum tests of fit. Annals of Mathematical Statistics, 1955, 26, 94-104.

T* 48. Tukey, J. W., A problem in the distribution of rankings. Memorandum Report No. 40, Statistical Research Group, Princeton University, Sept. 1949.

T 49. Tukey, J. W., Sums of random partitions of ranks. Annals of Mathematical Statistics,1957, 28, 987-992.

50. van der Vaart, H. R., On a basic distribution-free multidecision solution of a certain K-sample problem. Abstract from Proceedings International Mathematics Congress, Amsterdam, Sept. 1954.

51. Wallis, W. A., The correlation ratio for ranked data. Journal of the American Statistical Association, 1939, 34, 533-538.

T* 52. Whitney, D. R., A bivariate extension of the U statistic. Annals of Mathematical Statistics, 1951, 22, 274-282.

* 53. Wilcoxon, F., Some rapid approximate statistical procedures, American Cynamid Company pamphlet, 1949, 8-9.

54. Willerman, B., The adaptation and use of Kendall's coefficient of concordance (W) to sociometric-type rankings. Psychological Bulletin, 1955, 52, 132-133.

55. Wilson, E. B., and Worcester, Jane, The association of three attributes. Proceedings National Academy of Science, (U. S. A.), 1942, 28, 384-390.

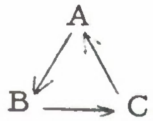
* 56. WILSON, K. V., A distribution-free test of analysis of variance hypotheses. Psychological Bulletin, 1956, 53, 96-101.

CHAPTER XIII

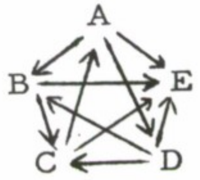
MISCELLANEOUS TESTS

The following chapter presents tests which do not appear to be readily categorizable within the topics covered by the previous chapters. They include tests for: transitivity of preference for a single judge, agreement among several judges, trend in location, trend in dispersion, goodness of fit, and peripheral association.

1. Paired Comparisons: "Consistency" of a Single Judge (Transitivity of Preference)

a. *Rationale.* Suppose that a judge is presented with each of the $\binom{n}{2}$ pairs of objects which can be made with n objects and is required to express a preference for one of the members of each pair over its paired mate. If his preferences are transitive and are based upon subjectively real differences, then for any three objects, say A, B and C, if A is preferred over B and B is preferred over C the judge must necessarily prefer A over C. Expressed differently, if the three objects are made the vertices of a triangle and if an arrow is placed between each pair of objects, pointing away from the preferred member of the pair, then if preferences are real and transitive the arrows will not all point in the same circular, i.e. clockwise, direction as is the case in the "inconsistent" triangle, or "circular triad", 

A test for transitivity, then, can be based upon whether or not the obtained number of circular triads is smaller than would be expected by chance. Let the n objects be placed at the vertices of an n sided polygon with arrows drawn between each pair of objects, indicating the direction of preference. There are $\binom{n}{2}$ pairs of objects and, therefore, $\binom{n}{2}$ arrows. Each arrow can have one of two directions. Therefore there are $2^{\binom{n}{2}}$ different patterns of arrow-directions which can be formed by changing directions of arrows in the polygon. The number of triads in the polygon is a constant, $\binom{n}{3}$; however, the number of circular triads depends upon the direction of the arrows. For each of the $2^{\binom{n}{2}}$ different patterns of arrow-directions there will



be some number of circular triads. Therefore the probability for that number or a smaller number of circular triads is simply the number of patterns of arrow-directions in which that number or a smaller number of circular triads occurs, divided by

$2^{\binom{n}{2}}$

b. Null Hypothesis. Each of the $2^{\binom{n}{2}}$ patterns of arrow-directions was equally likely to have been the one obtained. This will be the case if the judge actually has no real preferences in any of the $\binom{n}{2}$ choice situations and expresses preferences purely on a chance basis.

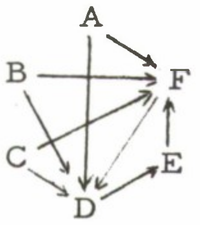
c. Assumptions. A preference is expressed for one of the members of each pair of objects, i.e., there are no tied choices. It is assumed that "trials" are randomly selected; this is necessary to insure that the sample of the judge's behavior is representative of his behavior in general. Random selection of judges is not assumed since inference is confined to the judge tested. Independence of choices is not assumed since a test for transitivity, in a sense, tests independence rather than assuming it.

d. Treatment of Ties. Ties should be obviated by using a forced choice technique. If they appear anyhow, the simplest procedure is probably to discard those objects for which the greatest number of ties exist and to continue the process until no ties exist among the remaining objects. The test may then be conducted upon the remaining number of objects.

e. Efficiency. No information available.

f. Application. Six vintages of a certain type wine are to be tested as to taste. The vintages are presented to a judge in pairs and he indicates the better tasting member of each pair. This is done for all 15 possible pairings with the following results, the arrow pointing away from the preferred member of each pair:

A → B, A → C, A → D, A → E, A → F, B → C, B → D, B → E, B → F, C → D, C → E, C → F, D → E, D ← F, E → F. Obviously the only intransitivity is D ← F, and only triads having DF as a side can be circular. The following polygon therefore shows only these triads. Only one of these triads, DEF, is circular.



Consulting Kendall's (30, 31, 32) tables it is found that, when $n = 6$, the probability of two or more circular triads is .949. Therefore the probability of one or less circular triads is 1 - .949 or .051. The .05 level of significance is not quite attained, therefore, and the hypothesis that preferences are either intransitive or determined by "chance" cannot be rejected in favor of the alternative hypothesis of "greater-than-chance" transitivity of preferences.

g. Discussion. Kendall apparently takes large values of d, the number of curcular triads, as his rejection region. Thus the test rejects the hypothesis of either chance or transitive preferences in favor of the alternative hypothesis that the judge's preferences are intransitive at a frequency so large that it would seldom occur by chance. However, one would expect this application to be somewhat less frequent than the one described.

h. Tables. The exact probability for d or more circular triads has been tabled (30, 31, 32) for cases in which $2 \leq n \leq 7$. When n is larger than 7, the probability of d or more circular triads is 1 minus the probability, read from chi-square tables, of

$$X^2 = \frac{2\binom{n}{3} - 8d + 4}{n-4} + \frac{n(n-1)(n-2)}{(n-4)^2} \quad \text{with} \quad \frac{n(n-1)(n-2)}{(n-4)^2}$$

degrees of freedom.

The probability of d-1 or fewer circular triads is 1 minus the probability, for d or more circular triads. It is therefore obtained by taking the complement of the probability given in the exact tables, or by taking the probability of chi-square as defined above, rather than its complement.

Counting the number of circular triads may prove difficult when n is not small. Kendall (30) has shown that a simpler method may be used to gain this information. An n x n table is constructed with each of the n objects being represented by one column and one row. If the i^{th} object is preferred over the j^{th} object, a 1 is entered in the cell of the i^{th} row and the j^{th} column; if the reverse is the case, a zero is entered. All cells, except those whose row and column represent the same object, are filled in. If the row totals are $a_1, a_2, \ldots, a_n$, then the number of circular triads, d, is given by $d = \frac{n(n-1)(n-2)}{6} - \frac{1}{2}\sum_{i=1}^{n} a_i(a_i-1)$.

i. Sources. 30, 31, 32, 40.

2. Paired Comparisons: Agreement among m Judges

a. Rationale. Suppose that each of m judges has expressed a preference for one of the members of a pair in each of the $\binom{n}{2}$ possible pairings of n objects and that it is desired to test whether or not the judges tend to agree among themselves. Let C_{ij} be the number of judges choosing object i over object j. Then the number of judges preferring j to i is $m-C_{ij}$. The C_{ij} judges preferring i to j can be paired with one another in $\binom{C_{ij}}{2}$ ways and each way represents an agreement between two judges that the i^{th} object is preferable to the j^{th}. Likewise there are $\binom{m-C_{ij}}{2}$ pairs of judges preferring j to i and there are that many "agreements" that j is preferable to i. The number of agreements as to the relative excellence of objects i and j, irrespective of which object is the one preferred, is therefore,

$\binom{C_{ij}}{2} + \binom{m-C_{ij}}{2}$. And the sum, $\sum \left[\binom{C_{ij}}{2} + \binom{m-C_{ij}}{2} \right]$, taken over all $\binom{n}{2}$ pairs of values of i and j (corresponding to pairings of objects with an object other than itself) is the total number of agreements among the m judges in all of the $\binom{n}{2}$ pairings of objects. This sum, represented by the symbol $\sum$, is the test statistic.

Now consider a table, such as that shown below, with m columns, corresponding to the m judges, and $\binom{n}{2}$ pairs of rows, each pair of rows corresponding to a pairing of objects and each row in a pair corresponding to preference for one of the two objects over its paired mate. In each cell of the table enter a 1 if the row object was preferred over the other object in the row pair by the column judge, otherwise enter a zero. There are two ways a judge can assign his preference to one of the members of a pair of objects, i.e., a 1 can be entered for either the first-listed or a second-listed object in a pair of rows. For each of these two ways, there are two ways in which a second judge can assign his preference to one of the members of that pair, etc., so since there are m judges there are 2^m ways in which their preferences can be assigned to the members of a single pair of objects. And since there are $\binom{n}{2}$ pairs of objects, there are $(2^m)^{\binom{n}{2}}$ or $2^{m\binom{n}{2}}$ ways in which m judges can assign their preferences among the members of n objects judged in pairs. Thus there are $2^{m\binom{n}{2}}$ different tables, i.e., tables with different patterns of cell entries, which can be formed by permuting 1s and 0s within their column and pair of rows. And if each judge assigns all his preferences randomly each of these $2^{m\binom{n}{2}}$ tables is equally likely. To each such table there corresponds a value of $\sum$ and, if preference assignments are random, the probability of this or a

		JUDGES					REQUIRED COMPUTATIONS		
		1	2	3	4	5	$\binom{C_{ij}}{2}$	$\binom{m-C_{ij}}{2}$	$\binom{C_{ij}}{2} + \binom{m-C_{ij}}{2}$
Object Pairings	1	1	1	0	1	1	6		6
	2	0	0	1	0	0		0	0
	1	1	0	1	0	1	3		3
	3	0	1	0	1	0		1	1
	1	0	1	1	0	0	1		1
	4	1	0	0	1	1		3	3
	2	1	1	1	1	1	10		10
	3	0	0	0	0	0		0	0
	2	0	0	1	0	0	0		0
	4	1	1	0	1	1		6	6
	3	0	0	0	0	0	0		0
	4	1	1	1	1	1		10	10

$\sum = 40$

larger value of $\sum$ is simply the number of the $2^{m\binom{n}{2}}$ tables giving rise to this or a larger value of $\sum$ divided by $2^{m\binom{n}{2}}$

b. Null Hypothesis. Each of the $2^{m\binom{n}{2}}$ tables is equally likely to have been the table actually obtained. This will be the case if each judge assigns his preferences randomly among the members of each pair of objects in which case agreements between judges will be accidental and the obtained number of such agreements will be determined by chance. See "Discussion".

c. Assumptions. There are no tied choices, i.e., in every choice situation one of the objects in a pair must be preferred over its mate. "Trials" are randomly selected; this assumption is necessary to insure that the sample of the judge's behavior is representative of his behavior in general. Random selection of judges is not assumed since inference is confined to the judges tested. Independence of choices is not assumed, rather it is tested.

d. Treatment of Ties. Ties should be obviated by using a forced choice technique. If they appear anyhow, the simplest procedure is probably to confine the test to those judges or to those objects for which no tied choices appear, making the necessary reductions in m and/or n.

e. Efficiency. No information available.

f. Application. Suppose that each of four judges compares three brands of chocolate ice cream in pairs and expresses a preference in each case. It is desired to test whether or not the judges tend to agree among themselves. Let the data be shown below:

		JUDGES				ROW	$\binom{\text{ROW TOTAL}}{2} = \binom{C_{ij}}{2}$ or $\binom{m-C_{ij}}{2}$
		I	II	III	IV	TOTAL	
Pairs of Brands	A	1	1	1	1	4	6
	B	0	0	0	0	0	0
	A	0	0	0	0	0	0
	C	1	1	1	1	4	6
	B	0	1	0	1	2	1
	C	1	0	1	0	2	1

The value of $\sum$ is 14. Entering Kendall's (30, 31) tables with $m = 4$, $n = 3$ and $\sum = 14$, we find that this value of $\sum$ has a probability of .043 of being equalled or exceeded. Therefore the null hypothesis of random assignment of preferences is rejected in favor of the alternative hypothesis that there is a nonchance degree of agreement among judges.

The probability obtained from tables could have been computed. There are $2^{m\binom{n}{2}}$ or 2^{12} possible patterns of preferences for the table shown. Greatest agreement would occur if in each pair of rows the ls were all in one row, the zeros in the other. There are two ways in which the 1s can all be in one row of a pair of rows, and since there are three pairs of rows there are $2^3 = 8$ ways in which the greatest agreement can occur, leading to a $\sum$ of 18. The next greatest amount of agreement occurs when in two pairs of rows the 1s are all in a single row of the pair and in the third pair of rows three 1s are in one row, the remaining 1 in the other row. There are 2^2 ways in which for both of two given pairs of rows all the 1s in a pair can be in one row. In the remaining pair of rows either row can be selected to contain the single 1, and the 1 can occur in any of its four cells, making eight ways of obtaining four 1s in one row and one 1 in the other row of a given pair. Finally, the pair of rows one of which contains three 1s, the other a single 1, could occur for any one of the three pairs of objects. Therefore there are $(2^2)(8)(3) = 96$ ways in which the next greatest $\sum$, 15, could be obtained. The next greatest amount of agreement is for the obtained case, where in two pairs of rows the 1s are all in one row of the pair, and in the remaining pair of rows each row contains two 1s. This case differs from the preceding one only in the number of ways of assigning 1s in the remaining pair of rows: there are $\binom{4}{2} = 6$ ways of placing two 1s in two of the four cells of a row. So for the last case there are $(2^2)(6)(3) = 72$ ways of obtaining a $\sum$ of 14. The probability for a $\sum$ of 14 or greater is therefore $\frac{8+96+72}{2^{12}} = .043$.

A somewhat simpler tabulational procedure than that given above is to form an n x n table with rows and columns both representing objects. The cell entry, r_{ij}, in the ith row and jth column is the number of judges who prefer i when it is compared with j. The value of $\sum$ is found by summing $\binom{r_{ij}}{2}$ over all n(n-1) cells in the table corresponding to preferences for the row object over a different column object. For the data just given the table would be:

	A	B	C
A	X	4	0
B	0	X	2
C	4	2	X

and $\sum$ would be $\binom{4}{2} + \binom{0}{2} + \binom{0}{2} + \binom{2}{2} + \binom{4}{2} + \binom{2}{2} = 6 + 0 + 0 + 1 + 6 + 1 = 14$ as before.

g. Discussion. Strictly speaking, the null hypothesis is that all preferences are assigned randomly since the use of an unweighted $2^{m\binom{n}{2}}$ as the denomimator of a probability fraction implies that this is so, i.e., since the tables for $\sum$ are based upon its chance distribution. Preferences, of course, can be assigned quite systematically without there being any substantial measure of agreement in the group of judges as a whole. For example half the judges may always prefer the "alphabetically higher" object of a pair and half may always prefer the "alphabetically lower". In cases such as this one, where there are systematic but opposing biasses among judges, the null hypothesis as stated is false, but

$\sum$ will not assume an extreme value calling for its rejection.

The null hypothesis is likely to be rejected if there are systematic but unopposing, biasses, but this condition amounts to "agreement" among judges. Therefore, since rejection of the null hypothesis can only be caused by chance (to the degree implied by the significance level used) or by agreement among judges, it can be regarded, as a practical matter, as stating simply that there is no nonchance degree of agreement among judges.

It is to be noted that agreement among judges does not imply transitivity of preference. For example in paired comparisons of three objects, there might be complete agreement in that all judges prefer A to B, B to C and C to A, which set of preferences forms a circular, or "inconsistent" triad. Nor does transitivity for each judge imply agreement among judges. When either agreement or transitivity is lacking, it would not be legitimate to rank the n objects from best to worst on the basis of preferences expressed in paired comparisons.

A test for agreement among judges is useful when the thing being measured is of a strictly subjective nature, such as the relative deliciousness of a variety of flavors. The paired comparison technique is useful when the n things being compared differ along so many dimensions or in such a complex way that they cannot properly be ranked from best to worst. The technique is also useful when judgments are strongly affected by such sequential factors as the immediately preceding trial, the number of preceding trials and the interval between trials. This type of situation arises, for example, in taste testing where the sensitivity of the taste buds depends upon the nature, number and duration of the preceding stimuli and upon the interval between the present and the preceding stimulus. In order to compare properly two taste stimuli, they must not be separated by intervening stimuli. The method of paired comparisons, therefore, is generally used.

The test described, originating with Kendall and Smith (32), appears to be the simplest and easiest to apply. However, exact tests for paired comparisons have also been devised and tabled by Bradley and his colleagues (1, 5, 6, 7, 8, 9, 58). A test somewhat analogous to that of Kendall and Smith has been outlined and tabled by Cartwright (10). However it is not connected with the method of paired comparisons. Instead it tests multijudge reliability when each of m judges assigns each of n objects to one of K categories.

h. Tables. Exact probabilities have been tabled (30, 31, 32) for $\sum$ for the cases $m = 3$, $2 \leq n \leq 8$; $m = 4$, $2 \leq n \leq 6$; $m = 5$, $2 \leq n \leq 5$ and $m = 6$, $2 \leq n \leq 4$. When m or n exceed these values, approximate probabilities for $\sum$ may be obtained by referring

$$X^2 = \frac{4}{m-2}\left[-\binom{n}{2}\binom{m}{2}\frac{m-3}{2(m-2)} + \sum\right], \text{ with } \binom{n}{2}\frac{m(m-1)}{(m-2)^2} \text{ degrees}$$

of freedom, to the probability tables for chi-square. A correction for continuity may be made by subtracting 1 from $\sum$.

i. Sources. 4, 23, 30, 31, 32, 40, (See also 1, 5, 6, 7, 8, 9, 19, 27, 58.)

3. The Difference-Sign Test for Trend

a. Rationale. Suppose that N observations have been made in sequence upon a continuously distributed variable and it is desired to test whether or not the variable's fluctuations contain a temporal trend. Let each observation (except the first) be subtracted from the observation immediately preceding it, and record only whether the difference is positive or negative. The number of algebraic signs of one kind recorded for the N-1 subtractions is the test statistic. If there is a trend, signs of one kind should predominate. Suppose that the N observations were ranked in order of size from 1 to N. If there were no trend, then each of the N! permutations of the integers from 1 to N would be equally likely to be the obtained sequence of ranked observations. Therefore, in the absence of trend, the probability of obtaining m or more minus difference-signs is simply the number of the N! permutations of integers from 1 to N which yield m or more minus differences, when each integer is subtracted from the one preceding it, divided by N!

b. Null Hypothesis. Each of the N! permutations of the N observations is equally likely to have been the sequence obtained. If a monotonic trend exists, this will not be the case.

c. Assumptions. The sampled population is continuously distributed, i.e., there are no tied observations. Sampling is random in the sense that the moment at which an observation is taken is selected without knowledge as to the magnitude the observation will have at that moment.

d. Treatment of Ties. A small number of tied observations are a practical problem only when they are adjacent in sequence. In this case, for a conservative test, give all zero differences the sign least conducive to rejection of the null hypothesis. To minimize tie error in the long run, arbitrarily give half the zero differences a plus sign, half a minus sign.

e. Efficiency. Against normal regression alternatives, the difference-sign test has an asymptotic relative efficiency of zero with respect to the regression coefficient test, as well as with respect to a half-dozen distribution-free tests (55). See Table I in the Introduction. It is superior in efficiency to the turning points test. An A.R.E. of zero does not, of course, mean that the test is useless. (See Introduction.)

The test has been found to be consistent, and its power has been investigated, in the case of normal regression alternatives (20, 56).

f. Application. Seven observations are taken in sequence and are as follows: 95, 88, 86, 81, 84, 77, 72. Starting with the second observation and subtracting each observation from the preceding one we have the following sequence of difference-signs: +, +, +, -, +, +. Entering Moore and Wallis' tables with N = 7 and m = 1, we obtain .048 as the probability of 1 or fewer differences of like sign. Therefore, the null hypothesis of no trend can be rejected at the two tailed .048 level, or if the null hypothesis was that there is either no trend or an upward trend it could be rejected at the one-tailed .024 level of significance.

g. Discussion. Moore and Wallis (39) and Stuart (56) have also considered tests for correlation between two series of observations. Stuart aligns the two sequences of difference-signs, one below the other, and takes as his test statistic the number of columns containing like difference signs. Moore and Wallis tabulate the frequency of occurrence of each of the four possible combinations of sign among the two entries in a column and analyze by means of a fourfold table. Unfortunately these tests appear to be strictly legitimate only if neither series contains a real trend, in which case the true correlation would be zero. (See 39 page 161.) For large samples they may be useful as approximate tests.

h. Tables. Exact two-tailed probabilities for the number of difference-signs of one sign have been tabled by Moore and Wallis (39) for all values of N between 2 and 11.

For larger values of N, the number, m, of minus difference-signs is approximately normally distributed with mean (N-1)/2 and variance (N+1)/12. Therefore approximate probabilities may be obtained by entering the normal tables with

$$Z = \frac{m - \frac{N-1}{2}}{\sqrt{\frac{N+1}{12}}} .$$

A correction for continuity can be introduced by reducing the absolute value of the numerator by $\frac{1}{2}$. The probability obtained will be one-tailed unless the tables give two-tailed probabilities.

i. Sources. 20, 36, 39, 54, 55, 56.

4. Records Tests for Trend in Location or Dispersion

a. Rationale. Let the r^{th} observation in a sequence of n observations be called an upper record if it is larger, and a lower record if it is smaller, than all of the r-1 preceding observations. (By definition the first observation is not a record value.) If there is no trend in the sampled variable, then each of the n! permutations of the n observations was equally likely to have been the sequence obtained, and any statistic based upon records should have a chance value. On the other hand, if there is a monotonic upward (downward) trend in location, then each observation has a greater-than-chance likelihood of being an upper (lower) record and a smaller-than-chance likelihood of being a lower (upper) record, and the difference, d, defined as the number of upper records minus the number of lower records should tend to assume extreme positive (negative) values. Likewise if there is a monotonic trend toward increasing (decreasing) dispersion, then each observation has a greater (smaller) than chance likelihood of being a record of either type, and the sum, s, defined as the number of upper records plus the number of lower records, should tend to assume an extremely large (small) value. The probability for a given value of d, or of s, is simply the proportion of the n! permutations of the integers from 1 to n which yield that value of the statistic.

b. Null Hypothesis. Each of the n! possible permutations of the n untied observations was equally likely to have been the sequence obtained in the sample.

c. Assumptions. The sampled population is continuously distributed, i.e., there are no tied observations. Sampling is random and independent.

d. Treatment of Ties. The authors recommend that ties be broken randomly, i.e., that one should "rank the tied observations according to a random permutation of their serial order." However, for a conservative test resolve ties in the manner least conducive to rejection of the null hypothesis.

e. Efficiency. As a test for randomness against normal regression alternatives, the d test has an asymptotic relative efficiency of zero with respect to the best parametric test based on the regression coefficient and with respect to some half-dozen distribution-free tests. It is more efficient than either the difference-sign test or the turning points test, both of which have zero A.R.E. with respect to the d test, (55). See Table I in Introduction.

The power of both the d test and the D test (see Discussion), at the .05 level, against the alternative that the sampled variable is normally distributed with constant variance but with a positive linear trend in the mean has been tabulated by Foster and Stuart (20) for various sample sizes and degrees of trend. These power functions were obtained empirically by means of a large sampling experiment.

The d test is consistent against the alternative that the form of the sampled population's distribution remains constant while a location parameter increases by equal increments along the sequence. The s test is consistent against an analogous alternative involving trend in dispersion only (20). The authors believe these consistency properties to apply also to the round-trip tests (see Discussion).

f. Application. A rat makes the following sequence of time scores in running a very difficult maze, 460, 457, 459, 455, 453, 451, and it is desired to test whether or not the rat is learning. There are four lower and no upper records. Entering Foster and Stuart's (20) tables the probability that d = -3 or a larger value is found to be .985, so the probability that d = -4 or less is .015 and the hypothesis of no learning is rejected.

Had the rats' scores been 455, 457, 456, 453, 450, 465, 447, 463, 475,444, 449 there would have been three upper and four lower records. The hypothesis that the rat's variability was increasing with time (possibly indicating the testing and rejection of false hypotheses by the rat) can be tested by entering Foster and Stuart's tables with n = 11 and s = 7. The probability that s does not exceed 6 is found to be .964, so the probability of an s of 7 or greater is .036 and the hypothesis of constant variability is rejected in favor of the hypothesis that it is increasing.

g. Discussion. The statistics d and s are asymptotically independent. Therefore when n is quite large a general test of the null hypothesis of randomness against alternatives of nonrandomness can be made by combining probabilities for d and for s, using the conventional methods for combining independent probabilities.

The number of upper records, when proceeding from the first to the last observation is not necessarily the number of upper nor necessarily the number of lower, records when proceeding in the opposite direction, and, in fact, is unlikely to be so. Therefore additional "information" is contained in the "round-trip" statistic $D = d - d'$ where d' is analogous to d but counted by proceeding from the last observation to first. No exact small sample tables for D are available; however, when n is large, D is approximately normally distr ibuted with mean of zero, so $\frac{D}{\sigma_D}$ may be treated as a normal deviate and probabilities obtained from normal tables. Unfortunately σ_D is not easy to obtain; however, a few approximate values have been tabled: Table 4 of (20) gives empirical values of σ_D corresponding to n's of 10, 25, 50, 75, 100, and 125 based upon a large sampling experiment. The D test was found to be considerably more powerful than the d test on the basis of a sampling experiment conducted by its authors.

h. Tables. Tables have been published (20) which give the exact probability that d does not exceed given values when $3 \leqq n \leqq 6$. Other tables (20) give the exact probability that s does not exceed given values for $3 \leq n \leq 15$.

When these tables do not apply, approximate tests may be performed by taking $\frac{d - \bar{d}}{\sigma_d}$ or $\frac{s - \bar{s}}{\sigma_s}$ as normal deviates and obtaining approximate probabilities by referring them to normal tables. The value of $\bar{d}$ is zero. The values of $\bar{s}$, and the standard errors, σ_s and σ_d are given in Table 3 of (20) for values of n from 10 to 100 in steps of 5.

i. Sources. 11, 20, 55.

5. The S_1 Sign Test for Trend

a. Rationale. A number of tests for trend use as the test statistic the number of difference-signs of one type resulting from a series of subtractions of subsequent from earlier observations. However, these tests are not equally efficient. If a real monotonic trend exists, then the farther apart two observations are in the sequence the greater their difference in size is likely to be and the greater is the likelihood that the sign of their difference will correspond to the direction of the trend. Therefore, Cox and Stuart subtract the N^{th} observation from the first, the (N-1)st from the 2nd, the (N-2)nd from the 3rd, etc., and weight each difference-sign by the distance between the observations giving rise to it. Thus if h_{ij} is defined to be 1 when the i^{th} observation is greater than the j^{th}, i.e., if their difference-sign is plus, and to be zero when the reverse is the case, then Cox and Stuart's test statistic is

$$S_1 = \sum_{k=1}^{\frac{N}{2}} (N - 2k + 1)\, h_{k,\ N - k + 1}.$$

This statistic is asymptotically normally distributed with mean $N^2/8$ and variance $N(N^2 - 1)/24$, thus providing a large-sample, approximate test of significance. N must always be made an even number. When there is an odd number of observations, the middle observation is dropped.

b. Null Hypothesis. Each of the N! permutations of the N observations was equally likely to have been the sequence obtained.

c. Assumptions. The sampled population is continuously distributed, i.e., there are no tied observations. Sampling is random and independent.

d. Treatment of Ties. A small number of ties does not create a practical problem unless a k^{th} observation is tied with an N - k + 1st observation. In this event, resolve ties in the manner least conducive to rejection of the null hypothesis, for a conservative test; or, to minimize error in the long run, give h a value of

$\frac{1}{2}$, i.e., half way between a zero, indicating a minus, and a 1, indicating a plus.

e. Efficiency. As a test for randomness against normal regression alternatives, the S_1 sign test for trend in location has asymptotic relative efficiency of .86 relative to the best parametric test based on the regression coefficient and has an A.R.E. of .87 relative to Kendall's rank correlation test, i.e. Mann's T test. It is more efficient than a number of other distribution-free tests for trend (12, 55). See Table I in Introduction.

f. Application. Let the observations be 50, 51, 52, 34, 54, 56, 55, 51, 20, 47, 42, 43, 44, 41, 28, 35, 39, 36, 30, 31, 29, 23, 25, 18, 21. There are 25 observations, so the middle observation, 44, is dropped, leaving N = 24. The differences are (50-21), (51-18), (52-25), (34-23), (54-29), (56-31), (55-30), (51-36), (20-39), (47-35), (42-28), (43-41), and the corresponding values of h are 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1. The corresponding weights of h, followed in parentheses by the value of h are 23(1), 21(1), 19(1), 17(1), 15(1), 13(1), 11(1), 9(1), 7(0), 5(1), 3(1), 1(1). Thus

$$S_1 = \sum_{k=1}^{N/2} (N - 2k + 1)\, h_{k,\ N-k+1} = 137(1) + 7(0) = 137.$$

If the null hypothesis is true, this value has a mean of approximately $N^2/8 = 24^2/8 = 72$ and a variance of $N(N^2-1)/24 = 24(24^2-1)/24 = 575$ and $\frac{137 - 72}{\sqrt{575}} = 2.71$ is approximately a normal deviate. Entering the normal tables with this value we find that the obtained value of S_1 is significant at the two-tailed .01 level (or at the one-tailed .005 level). In view of the small value of N used, these probabilities should be regarded as very approximate.

g. Discussion. In addition to its use as a test for trend in location, the outlined technique, slightly modified, can also be used to test for trend in dispersion. (In which case it has A.R.E. of .74, under "parametric" conditions, relative to the maximum likelihood test.) The sequence of N observations is divided into r blocks, each block containing the same number, k, of consecutive observations, the N-rk "extra" observations being randomly selected and discarded. The range, w, is determined for each block. The sequence of ranges of consecutive blocks is then treated as a sequence of r "observations" and tested for trend in location by means of the S_1 sign test already outlined. A monotonic trend in "location" of ranges is equivalent to a monotonic trend in the dispersion of the observations upon which the ranges are based.

h. Tables. There appear to be no tables of exact probabilities, so the test should not be used when N is small. As N approaches infinity, the distribution of

$$\frac{S_1 - \frac{N^2}{8}}{\sqrt{\frac{N(N^2-1)}{24}}}$$

approaches the normal distribution whose mean is zero and whose variance is unity. Therefore tables of the normal distribution can be used to obtain approximate probabilities when N is moderately large.

i. Sources. 12, 55.

6. David's Combinatorial Tests of Fit

a. Rationale. Suppose that an experimenter has a sample of N observations and that he wishes to test whether or not the sample came from an hypothesized population whose distribution he can specify completely. Since the population distribution, under the null hypothesis, is known, it can be divided into N nonoverlapping vertical strips, each of which contains the same area, 1/N.

To these N vertical strips there will be N corresponding ranges of abscissa values, each range having equal probability, 1/N, of "containing" an observation drawn randomly from the population. Now let N observations be drawn and let Z be the number of ranges containing zero observations, i. e. no observations. If the sample was drawn from the hypothesized population, then Z will have a "chance" value; if the sample was actually drawn from some other population, then Z will tend to assume large values with greater-than-chance probability. The probability of a given value of Z, when the null hypothesis is true, is simply the number of ways N balls can be dropped into N boxes or compartments so as to leave an unspecified Z compartments empty, divided by the number of ways N balls can be dropped into N boxes without restriction. These probabilities have been tabled by David (16).

The above test is a test of fit against general alternatives. However, if the experimenter suspects certain alternatives to be more likely than others, he may wish to specify the general location of the "empty compartments", i. e., ranges containing no sample observation. For example, if the true population is believed to have the same form as the hypothesized population but a larger median, then one would expect more empty compartments below the median of the hypothesized population than above it. Likewise, if the true and hypothesized populations are symmetrical and have equal means but different variances, the variance of the true population being the larger, then one would expect more empty compartments in the middle than at the extremes of the hypothesized distribution. David (16), therefore, has proposed a second test in which the hypothesized distribution is divided into 2N nonoverlapping vertical strips of equal area, of which N are selected to be the "test" compartments. A sample of N observations is then drawn and the number, Z, of empty compartments among the predesignated N test compartments is counted. Probabilities for Z in this second test have also been tabled by its author.

b. Null Hypothesis. Each of the N sample observations was equally likely to have been drawn from each of the N (or in the case of the second test, 2N) ranges of abscissa values corresponding to equal areas of the hypothesized distribution. This will be the case if the hypothesized distribution is the distribution sampled and if all assumptions are met.

c. Assumptions. Sampling is random and independent and there is zero probability that an observation will be tied for inclusion in adjacent abscissa ranges, i.e., that it will fall at the common endpoint of two abscissa ranges.

d. Treatment of Ties. If the hypothesized distribution is discontinuous and an endpoint of an abscissa range happens to coincide with one of the discrete population values, the test had best be avoided. Ties due to this cause could, of course, be broken and assigned among the two adjacent ranges in the same proportion as would be required to break the relative frequency of the discrete value in order to maintain equal areas in the hypothesized distribution.

If the hypothesized distribution is continuous, ties may be broken by assigning them in the manner least conducive to rejection of the null hypothesis, by assigning half of each group of ties to each of the two ranges tied for, or by breaking them randomly. See Introduction.

e. Efficiency. Formulae by which to obtain power functions are given (16) for both tests by their author, and certain power comparisons are made. The power of the first test and the power of chi-square to reject the hypothesis that the population is normally distributed with zero mean and unit standard deviation was obtained (using $N = 30$ and $\alpha = .05$) when the distribution and mean are as hypothesized but the standard deviation is 4/3. The ratio of the power of the zeros test to that of chi-square was .968.

f. Application. It is hypothesized that a certain population is normally distributed with a mean of 500 and a standard deviation of 10. In order to test this hypothesis, a sample of six observations is drawn from the population in question, their values being: 457, 462, 489, 515, 538, 564. From normal tables we find that

$\frac{1}{3}$ of the area of a normal curve is between $\pm .4307\sigma$ of the mean and

$\frac{2}{3}$ of the area, between $\pm .9674\sigma$ of the mean. Therefore since it

is symmetrical, the normal curve is divided into six equal and non-overlapping areas by the points $\mu - .9674\sigma$, $\mu - .4307\sigma$, μ, $\mu + .4307\sigma$, and $\mu + .9674\sigma$. Substituting 500 for μ and 10 for σ, these points become: 490.326, 495.693, 500.000, 504.307, and 509.674, and the six ranges or "compartments" are $-\infty$ to 490.326, 490.326 to 495.693, 495.693 to 500.000, 500.000 to 504.307, 504.307 to 509.674, and 509.674 to $+\infty$. The six sample observations all fall into two of the six ranges leaving four compartments empty. Entering David's tables with $N = 6$ and $Z = 4$, the probability of four or more empty compartments is found to be .0200 and the null hypothesis is therefore rejected at better than the .05 level of significance.

This probability could have been computed. The four empty compartments could have been selected in $\binom{6}{4}$ or 15 ways. The six observations can occupy the remaining two boxes in the following ways, the denominator of the multinomial expression in each case indicating the split of the six observations between the two compartments: $\frac{6!}{1!\,5!} + \frac{6!}{2!\,4!} + \frac{6!}{3!\,3!} + \frac{6!}{4!\,2!} + \frac{6!}{5!\,1!} = 6 + 15 + 20 + 15 + 6 = 62$. Finally, there are $N^N = 6^6 =$ 46,656 ways in which six observations can be assigned to six compartments without restriction as to how many are to be empty. The probability of exactly four empty compartments is therefore $\frac{15(62)}{46,656}$ or .0199. By similar reasoning, the probability of exactly five empty compartments is $\frac{\binom{6}{5}\frac{6!}{6!}}{46,656} = \frac{6}{46,656} = .0001$. So the probability of four or more empty compartments is $.0199 + .0001 = .0200$. The general formula for exactly Z empty compartments when there are N observations and N compartments is

$$\frac{\binom{N}{Z}}{N^N} \sum \frac{N!}{t_1!\, t_2! \,\ldots\, t_{N-Z}!}$$

where the summation is taken over all values of $t_1, t_2, \ldots, t_{N-Z}$ such that none of the N-Z t's is zero and such that the sum of the t's is N.

In the example given the alternative hypothesis was a general one. Had the experimenter suspected that the true and hypothesized populations would differ mainly in variance, if they differed at all, David's second test would be more appropriate. In this case, the hypothesized distribution would have been divided into 12 equal areas and the central six might have been chosen as test compartments if the experimenter suspected that the true distribution had a greater-than-hypothesized variance. None of the six sample observations fell into any of these six compartments, so Z would have been 6. Entering David's tables for her second test with $Z = 6$ and $2N = 12$, a somewhat lower probability of .0156 is found, as would be expected for a test making use of additional "information". The second test, of course, can be significant for either small or large values of Z or for both, depending upon the alternative hypothesis and whether or not it is two tailed.

g. Discussion. It is to be noted that the hypothesized distribution must be completely known prior to sampling. None of its parameters should be estimated from the sample. If this stricture is observed, then each of the N sample observations was equally likely to have been drawn from each of the N abscissa ranges, as required by the null hypothesis, and all N observations could have been drawn from any specified set of ranges or compartments. However, suppose that the distribution median is to be estimated from the sample median. Then the sample observations cannot possibly all have been drawn from the $\frac{N-1}{2}$ leftmost or from the $\frac{N-1}{2}$ rightmost compartments of the distribution whose median is the same as their own. It is clear, therefore, that the mathematical model upon which the test is based requires that the hypothesized population distribution be completely known in advance of sampling. To facilitate division into equal areas, it is also desirable that the distribution be extensively tabled.

h. Tables. Exact point probabilities for Z as well as probabilities cumulated to approximately the .05 level of significance, have been tabled (16) for $3 \leqq N \leqq 20$ for the first test in which Z is the number of unoccupied abscissa ranges each of which had probability $\frac{1}{N}$ of containing any given one of the N sample observations. For values of N greater than 20, Z is approximately normally distributed with mean and variance given in (16), and its probabilities may be obtained by referring the critical ratio to normal tables; however, the calculations are laborious. For values of $N \geqq 30$, the author suggests that the hypothesized distribution be divided into six or more equal areas, such that N divided by the number of areas yields an expected frequency of five or more for all compartments, and that the usual chi-square test of fit be applied.

Exact point probabilities for Z, and probabilities cumulated to approximately the .05 level of significance, have been tabled (16) for $1 \leqq N \leqq 10$ for the second test in which the hypothesized distribution is divided into 2N equal areas, N of which are selected as "test compartments", and in which Z is the number of these test compartments which are unoccupied by any of the N observations in the sample. Again, when N exceeds 10 Z is approximately normally distributed with mean and variance given by David, so probabilities can be obtained by forming the critical ratio and referring it to normal tables.

i. Sources. 16.

7. The Quadrant Sum (or "Corner") Test for Peripheral Association

a. Rationale. Suppose that an X measurement and a Y measurement have been taken on each of 2n objects and that it is desired to test whether or not X and Y are correlated. Let the 2n points be plotted as a scattergram and let a vertical line be drawn through the sample X-median and a horizontal line through the sample Y-median. Now find the rightmost point in the scattergram and, proceeding toward the middle of the scattergram, count the number of points passed before the Y median must be crossed to pick up the next point. This is the largest value for the number of rightmost X-values, all of which lie on one side of the Y-median. Call this number R_A if the points are all above the Y median and R_B if they are all below it. Next find the leftmost point and proceed analogously, calling the L leftmost points on one side of the Y median L_A if they are all above the median and L_B if they are all below it. Now find the uppermost point and proceed downward counting the number of points until the X median must be crossed to obtain the next point. Let this number of points be A_R if they are all on the right side of the X-median and A_L if they are all on the left. Finally, find the lowest point and proceed upward and analogously, calling the number of points B_R if they are all to the right, and B_L if they are all to the left, of the X median.

If the X and Y variables are correlated, the scattergram points should tend to lie in one pair of the diagonal quadrants formed by the lines through the X and Y medians. R_A, A_R, L_B and B_L all refer to points in the upper right or lower left quadrants and are therefore given a positive sign. Likewise, L_A, A_L, R_B and B_R refer to points in the upper left or lower right quadrants and therefore are given a minus sign. The four numbers actually recorded, each preceded by the proper algebraic sign, therefore yield an algebraic sum which can be used as the test statistic. Consider the X values to have been ranked from 1 to 2n and the Y values likewise to have been ranked from 1 to 2n and recorded below the X ranks. There are (2n)! ways in which the Y ranks can be permuted, and each way represents a different set of pairings or assignments of Y values to X values. The probability of a given quadrant sum or one more extreme is therefore the number of these (2n)! possible sets of assignments of Y values to X values which yield the given, or more extreme, quadrant sum, divided by (2n)!, the number of possible assignments.

b. Null Hypothesis. Each of the (2n)! possible sets of 2n pairs of X and Y values, which can be made with the obtained values of X and Y, was equally likely to have been the set obtained as a sample.

c. Assumptions. Sampling is random and independent and the sampled population is continuously distributed, i.e., there are no tied values.

d. Treatment of Ties. Tied observations create a practical problem when they occur at the "crossover point", i.e., when the manner in which the tie is broken affects the number of extrememost points. When this occurs, the authors suggest dividing the number of points in the tied group which are on the same side of the median as the more extreme points, by one plus the number of points in the tied group which are on the opposite side of the median, and counting the result as the number of "extrememost" points in the tied group. They regard this procedure as a conservative one. A more conservative technique would be to resolve all ties (including extreme observations lying on the X or Y median) in whatever manner is least conducive to rejection of the null hypothesis.

e. Efficiency. No information available.

f. Application. Consider the following data, in which the points are arranged in order of increasing X-value: (15, 71), (21, 68), (23, 75), (28, 63), (30, 57), (33, 59), (44, 65), (46, 66), (49, 52), (55, 48). The X median lies between 30 and 33, and the Y median between 63 and 65. The rightmost point, i.e., the largest X value is 55 and proceeding inward two points (55, 48) and (49, 52) are counted before a point is reached whose Y value is on the other side of the Y median. Thus $R_B = 2$ and $R_A = 0$. Likewise L_A is found to be 3 (so L_B is zero), since the three points (15, 71), (21, 68) and (23, 75) with lowest X values all have Y values above the Y median while the Y value paired with the fourth largest X value is below the Y median. The point with largest Y value is (23, 75) and the points (15, 71) and (21, 68) have diminishingly extreme Y values and X values on the same side of the X median, while the fourth largest Y value, 66, is paired with an X of 46 which is on the opposite side of the X median from the preceding points. Therefore A_L is three and A_R is zero. The points in order of increasing Y value are (55, 48), (49, 52), (30, 57) . . . of which the first two have X values to the right of the X median, the third having an X value to the left of the median. Therefore $B_R = 2$ and $B_L = 0$. Giving these values the algebraic signs corresponding to the associated quadrant, the quadrant sum is $+(R_A + A_R + L_B + B_L) - (L_A + A_L + R_B + B_R) = +(0 + 0 + 0 + 0) - (3 + 3 + 2 + 2) = -10$. Entering Olmstead and Tukey's tables (48), with $2n = 10$ we find the probability of a quadrant sum equal to or more extreme than 10 in absolute value to be .0642. A null hypothesis of no association could not be rejected at the two-tailed .05 level. However if the null hypothesis were that there is either no association or a positive correlation, it could be rejected at the one-tailed .05 level in favor of the alternative hypothesis of negative correlation.

g. Discussion. The quadrant sum or "corner" test for association obviously is especially sensitive to correlation between values at their extremes, at least at the extremes of one of the two variables. It tends, however, to ignore correlation within the central portion of the scattergram. Therefore, while providing an excellent test for "peripheral" association, it is, as its authors point out, of "unknown usefulness" "when uniform attention to the whole scatter diagram is desired."

The test can be extended to test for association between more than two variables, and its authors have provided a small table of probabilities for the "octant sum", the test statistic in the case of three variables.

The four nonzero numbers which make up the quadrant sum are not independent since a single point can be counted twice, i.e., an extreme-right point above the Y median may also be an extreme-high point to the right of the X median and be counted in both R_A and A_R. This lack of independence could be avoided by counting R_A, R_B, L_A and L_B first, then discarding these points before counting A_R, A_L, B_R and B_L (or ignoring them in the counting process). If this were done the "quadrant sum" so obtained would have a slightly different distribution than the quadrant sum defined by Olmstead and Tukey. However, its probabilities could now be obtained by the methods outlined in the chapter on exceedances. The n points above the Y median may be regarded as the first sample, the n points below it as the second. Let X_l and X_r be the most extreme leftward and rightward points above the Y median which are not counted in L_A or R_A. Then exceedance formulae can be used to determine the a priori probability that in the second sample, i.e. below the median, exactly L_B points will have X values smaller than X_l and exactly R_B will have X values exceeding X_r. The $L_A + L_B + R_A + R_B$ points can then be discarded and an analogous procedure can be applied to the two "samples" consisting of the h points to the left of the X median and the $2n - L_A - L_B - R_A - R_B - h$ points to the right of it. Since the X and Y values are independent under the null hypothesis, the two probabilities can be combined to obtain an overall probability for a quadrant sum whose components are exactly R_A, R_B, L_A, L_B, A_R, A_L, B_R and B_L, the appropriate algebraic signs, of course, being added to obtain the quadrant sum. Tabulation or calculation of probabilities in this manner would be quite tedious, largely because the same quadrant sum can be obtained in a variety of ways, depending on the values of the eight components. The method has been outlined primarily to show the nearness of relationship of the quadrant sum test to exceedances theory.

h. Tables. Exact two-tailed probabilities for a quadrant sum equal to or greater than k have been tabled (48) for the cases 2n = 2, 4, 6, 8, 10 and 14 with asymptotic probabilities for 2n = infinity. These tables should suffice in most cases since the probabilities at $2n = 14$ are very close to those for $2n = $ infinity, excepting those probabilities smaller than .01. In fact the probability of a given quadrant sum is so insensitive to sample size that the authors have presented a table of approximate probabilities for the quadrant sum which does not use n as a parameter; it is merely stipulated that the table is inapplicable if the absolute value of the quadrant sum equals or exceeds $2(2n) - 6$.

i. Sources. 48.

8. Additional Tests

Mood (38) has proposed a rank test for dispersion which has asymptotic relative efficiency of .76 relative to the F test when both tests are two-tailed and .87 when they are both one-tailed. If there are m X-observations and n Y-observations from continuous distributions, the observations are ranked from 1 to m+n irrespective of sample. The test statistic, W, is the sum of the squared deviations of Y ranks from the average rank of all observations, i.e., $W = \sum_{i=1}^{n} (r_i - \frac{m+n+1}{2})^2$, where r_i is the rank of the i^{th} Y observation. Since W can assume large values due to differences in either location or dispersion, it must be assumed that the X and Y populations have identical location parameters. The probability of W under the null hypothesis is simply the proportion of the $\binom{m+n}{m}$ ways of obtaining m X-ranks and n Y-ranks from m+n ranks, which give a calculated value of W equal to or greater than that obtained. Unfortunately these probabilities do not appear to have been tabled; however, under the null hypothesis, W has a mean of $n(m+n+1)(m+n-1)/12$ and a variance of $mn(m+n+1)(m+n+2)(m+n-2)/180$ and is asymptotically normally distributed.

Another rank test for dispersion has been proposed by Lehmann (33) and developed further by Sundrum (57). Let m

X-observations and n Y-observations be drawn from continuous distributions and ranked from 1 to m+n. Then form each of the $\binom{m}{2}$ possible pairs of X-ranks and each of the $\binom{n}{2}$ possible pairs of Y-ranks. Finally, form each of the $\binom{m}{2}\binom{n}{2}$ quadruples of possible pairs of X-ranks paired with possible pairs of Y ranks, and count the number of quadruples, Q, in which both X-ranks are either greater or smaller than both Y-ranks. This number can be obtained from a formula given by Lehmann. The probability that the number of such quadruples will be Q or greater, if the null hypothesis of identical populations is true, is simply the proportion of the $\binom{m+n}{m}$ divisions of m+n ranks into m and n ranks which yield that value of Q or a larger one. The test is consistent, (if the sampled populations are continuous and if ties are randomized) but not unbiassed. Sundrum defines a statistic

$$L = \frac{Q}{\binom{m}{2}\binom{n}{2}}$$

and has tabled some of its probabilities.

A second "quadruple" test for dispersion suggested by Lehmann (33, page 169) appears not to be entirely distribution-free (38, page 521).

A test for dispersion, somewhat similar to Rosenbaum's (see Exceedances), has been published by Kamat (28). Two samples are drawn from populations assumed to be continuously distributed and to have the same location parameters. The n X-observations and m Y-observations, defined so that $m \geqq n$, are ranked from 1 to m+n, in order of magnitude, irrespective of sample. The test statistic is then $D_{n,m} = R_n - R_m + m$ where R_n and R_m are the ranges of the ranks of the X and Y observations respectively. By applying the Method of Randomization to the $\binom{m+n}{n}$ ways of assigning n ranks to Xs and m to Ys, exact probabilities can be obtained for $D_{n,m}$. Probabilities have been tabled for values of $m+n \leqq 20$, and a method is outlined by which to obtain approximate probabilities when $m+n > 20$.

If a sample of size K is drawn from a population consisting of the ranks from 1 to N, the sample mean, $\bar{r}$, will be approximately normally distributed with mean $\frac{N+1}{2}$ and variance $\sigma_{\bar{r}}^2 = \frac{N^2 - 1}{12K} (1 - \frac{K}{N})$ if K is large. Therefore, Locks (34) refers the critical ratio, $\frac{\bar{r} - \frac{N+1}{2}}{\sigma_{\bar{r}}}$ to normal tables to test whether or not a random sample has been drawn from the hypothesized population. He also uses the statistic chi-square $= \frac{12K \sum (r - \bar{r})^2}{(K-1)(N^2-1)}$ with K-1 degrees of freedom to test whether sample variance and population variance are comparable.

If the hypothesized distribution is completely known and tabled and is continuous, then goodness of fit can be tested by methods using the probability integral transformation (15, 17, 44, 49, 50, 51). If a sample of N observations is drawn from the hypothesized population, each sample observation's a priori probability may be obtained from tables of cumulative probabilities for the hypothesized distribution. Each observation may therefore be regarded as an independent test of significance and the overall probability for the N tests may be obtained by the usual methods of combining probabilities of independent tests of significance. If random sampling is assumed the hypothesis that the observations were drawn from a completely specified distribution may be tested. Conversely if it is assumed, i.e. known, that the observations came from a specified distribution, the randomness of sampling may be tested. In either case, no population parameters should be estimated from the sample; they must be specified in advance of sampling. Extensive tables exist for a test statistic, P_{λ_n}, based on this method.

An interesting and very simple method of linear curve fitting has been described by Nair and his colleagues (45, 46). One calculates the X mean and Y mean for the smallest 1/3 of the observations and finds the point for which they are abscissa and ordinate; he then does likewise for the largest 1/3 of the observations and draws a straight line through these two points.

Further distribution-free tests and methods are merely listed in the bibliography. Some are exact and provide tables of probabilities, but lack simplicity either conceptually or in application. Others are approximations for which there corresponds no exact small-sample probability formula.

BIBLIOGRAPHY

1. Abelson, R. M. and Bradley, R. A., A 2 X 2 factorial with paired comparisons. Biometrics, 1954, 10, 487-502.

2. Bartholomew, D. J., A sequential test of randomness for events occurring in time or space. Biometrika, 1956, 43, 64-78.

3. Barton, D. E. and David, F. N., Tests for randomness of points on a line. Biometrika, 1956, 43, 104-112.

4. Bose, R. C., Paired comparison designs for testing concordance between judges. Biometrika, 1956, 43, 113-121.

5. Bradley, R. A., Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparisons. Biometrics, 1954, 10, 375-390.

6. Bradley, R. A., Some large-sample results on power and power comparisons, Undated mimeographed report, Virginia Agricultural Experiment Station, Virginia Polytechnic Institute, Blacksburg, Virginia.

7. Bradley, R. A., Some statistical methods in taste testing and quality control. Biometrics, 1953, 9, 22-38.

T 8. Bradley, R. A., The rank analysis of incomplete block designs II. Additional tables for the method of paired comparisons. Biometrika, 1954, 41, 502-537.

T* 9. Bradley, R. A. and Terry, M. E., Rank analysis of incomplete block designs I. The method of paired comparisons. Biometrika, 1952, 39, 324-345.

T* 10. Cartwright, D. S., A rapid non-parametric estimate of multi-judge reliability. Psychometrika, 1956, 21, 17-29.

T 11. Chandler, K. N., The distribution and frequency of record values. Journal of the Royal Statistical Society (B), 1952, 14, 220-228.

* 12. COX, D. R. and STUART, A., Some quick sign tests for trend in location and dispersion. Biometrika, 1955, 42, 80-95.

* 13. Daniels, H. E., A distribution-free test for regression parameters. Annals of Mathematical Statistics, 1954, 25, 499-513.

14. Darling, D. A., On a class of problems related to the random division of an interval. Annals of Mathematical Statistics, 1953, 24, 239-253.

T 15. David, F. N., On the P_{λ_n} test for randomness: remarks, further illustration, and table of P_{λ_n} for given values of $-\log_{10}\lambda_n$. Biometrika, 1934, 26, 1-11.

TT** 16. DAVID, F. N., Two combinatorial tests of whether a sample has come from a given population. Biometrika, 1950, 37, 97-110.

17. David, F. N. and Johnson, N. L., The probability integral transformation when parameters are estimated from the sample. Biometrika, 1948, 35, 182-190.

18. Domb, C., The problem of random intervals on a line. Proceedings Cambridge Philosophical Society, 1947, 43, 329-341.

19. Dykstra, O., A note on the rank analysis of incomplete block designs - applications beyond the scope of existing tables. Biometrics, 1956, 12, 301-306.

TT** 20. FOSTER, F. G. and STUART, A., Distribution-free tests in time-series based on the breaking of records. Journal of the Royal Statistical Society, (B), 1954, 16, 1-22.

21. Gardner, A., Greenwood's 'problem of intervals': An exact solution for N = 3. Journal of the Royal Statistical Society (B), 1952, 14, 135-139.

22. Greenwood, M., The statistical study of infectious diseases. Journal of the Royal Statistical Society, 1946, 109, 85-103.

23. van der Heiden, J. A., On a correction term in the method of paired comparisons. *Biometrika*, 1952, 39, 211-212.

T* 24. Hodges, J. L., A bivariate sign test. *Annals of Mathematical Statistics*, 1955, 26, 523-527.

T* 25. Hoeffding, W., A non-parametric test of independence. *Annals of Mathematical Statistics*, 1948, 19, 546-557.

26. Housner, G. W. and Brennan, J. F., The estimation of linear trends. *Annals of Mathematical Statistics*, 1948, 19, 380-388.

27. Jackson, J. E. and Fleckenstein, Mary, An evaluation of some statistical techniques used in the analysis of paired comparison data. *Biometrics*, 1957, 13, 51-64.

T* 28. Kamat, A. R., A two-sample distribution-free test. *Biometrika*, 1956, 43, 377-387.

29. Katz, L., The distribution of the number of isolates in a social group. *Annals of Mathematical Statistics*, 1952, 23, 271-276.

TT** 30. KENDALL, M. G., *Rank correlation methods*, 2nd Ed., New York: Hafner, 1955.

TT 31. Kendall, M. G., *The advanced theory of statistics. Vol. I*, London: Griffin, 1947, 421-435.

T* 32. Kendall, M. G. and Smith, B. B., On the method of paired comparisons. *Biometrika*, 1939, 31, 324-345.

** 33. Lehmann, E. L., Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics*, 1951, 22, 165-179.

** 34. Locks, M. O., Two nonparametric tests for testing the randomness of samples drawn from finite populations. Abstract of paper given at Oklahoma Adademy of Science, *Froceedings of the Oklahoma Academy of Science*, 1953, 34, 195-196.

35. Mack, C., An exact formula for $Q_k(n)$, the probable number of k-aggregates in a random distribution of n points. *Philosophical Magazine*, 1948, 39, 778-790.

36. Mann, H. B., On a test of randomness based on signs of differences. Annals of Mathematical Statistics, 1945, 16, 193-199.

37. Mauldon, J. G., Random division of an interval. Proceedings of the Cambridge Philosophical Society, 1951, 47, 331-336.

* 38. Mood, A. M., On the asymptotic efficiency of certain non-parametric two-sample tests. Annals of Mathematical Statistics, 1954, 25, 514-522.

T* 39. MOORE, G. H. and WALLIS, W. A., Time series significance tests based on signs of differences. Journal of the American Statistical Association, 1943, 38, 153-164.

40. Moran, P. A. P., On the method of paired comparisons. Biometrika, 1947, 34, 363-365.

41. Moran, P. A. P., The random division of an interval. Journal of the Royal Statistical Society, (B), 1947, 9, 92-98.

42. Moran, P. A. P., The random division of an interval, II. Journal of the Royal Statistical Society, (B), 1951, 13, 147-150.

43. Mullemeister, H., Mean lengths of line segments. American Mathematical Monthly, 1945, 52, 250-252.

44. Nair, K. R., A note on the exact distribution of λ_n. Sankhyā, 1937, 3, 171-174.

45. Nair, K. R. and Banerjee, K. S., A note on fitting of straight lines if both variables are subject to error. Sankhyā, 1942, 6, 331.

46. Nair, K. R. and Shrivastava, M. P., On a simple method of curve fitting. Sankhyā, 1942, 6, 121-132.

47. Noether, G. E., Sequential tests of randomness, Report No. OSR-TN-54-65, Boston University report under contract AF 18(600)-778.

T** 48. OLMSTEAD, P. S. and TUKEY, J. W., A corner test for association. Annals of Mathematical Statistics, 1947, 18, 495-513. (Also entitled "Testing peripheral association", Bell Telephone System Monograph No. B-1515.)

49. Pearson, E. S., The probability integral transformation for testing goodness of fit and combining independent tests of significance. Biometrika, 30, 134-148.

* 50. Pearson, K., On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 1933, 25, 379-410.

* 51. Pearson, K., On a new method of determining 'goodness of fit'. Biometrika, 1934, 26, 425-442.

* 52. Sherman, B., A random variable related to the spacing of sample values. Annals of Mathematical Statistics, 1950, 21, 339-361.

53. Silberstein, L., The probable number of aggregates in random distributions of points. Philosophical Magazine, 1945, 36, 319-336.

54. Stuart, A., Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. Journal of the American Statistical Association, 1954, 49, 147-157.

55. Stuart, A., The efficiencies of tests of randomness against normal regression. Journal of the American Statistical Association, 1956, 51, 285-287.

56. Stuart, A., The power of two difference-sign tests. Journal of the American Statistical Association, 1952, 47, 416-424.

T 57. Sundrum, R. M., On Lehmann's two-sample test. Annals of Mathematical Statistics, 1954, 25, 139-145.

58. Terry, M. E., Bradley, R. A. and Davis, L. L., New designs and techniques for organoleptic testing. Food Technology, 1952, 6, 250-254.

59. Walsh, J. E., Correction to "Some nonparametric tests of whether the largest observations of a set are too large or too small". Annals of Mathematical Statistics, 1953, 24, 134-135.

60. Walsh, J. E., Some estimates and tests based on the r smallest values in a sample. Annals of Mathematical Statistics, 1950, 21, 386-397.

T 61. Walsh, J. E., Some nonparametric tests for Student's hypothesis in experimental designs. Journal of the American Statistical Association, 1952, 47, 401-415.

62. Walsh, J. E., Some nonparametric tests of whether the largest observations of a set are too large or too small. Annals of Mathematical Statistics, 1950, 21, 583-592.

CHAPTER XIV

TCHEBYCHEFF INEQUALITIES

In 1853 Bienaymé discovered, and in 1867 Tchebycheff rediscovered, a mathematical inequality variously called the Bienaymé-Tchebycheff inequality, or, more frequently, simply Tchebycheff's inequality. Following the essentials of a derivation presented by Hoel (10), let f(x) be a continuous distribution function with finite variance and mean u. Then by definition $\sigma^2 = \int_{-\infty}^{+\infty} (x-u)^2 f(x)dx$. This integral can be divided into three components whose sum it equals. Thus

$$\sigma^2 = \int_{-\infty}^{u-k\sigma} (x-u)^2 f(x)\,dx + \int_{u-k\sigma}^{u+k\sigma} (x-u)^2 f(x)\,dx + \int_{u+k\sigma}^{+\infty} (x-u)^2 f(x)\,dx$$

The second integral must be positive if k is positive; therefore, if $k \geq 0$, dropping the second integral must either diminish the value of the right-hand side of the equation or else leave it unaffected. Thus

$$\sigma^2 \geq \int_{-\infty}^{u-k\sigma} (x-u)^2 f(x)\,dx + \int_{u+k\sigma}^{+\infty} (x-u)^2 f(x)\,dx.$$

For the first integral, of all the values of x between $-\infty$ and $u-k\sigma$, that which will make $(x-u)^2$ smallest is that which is closest to u, namely $u-k\sigma$. Similarly, for the second integral that possible value of x which minimizes $(x-u)^2$ is $u+k\sigma$. The inequality must hold therefore if these values are substituted for x in the coefficient

$(x-u)^2$. Therefore

$$\sigma^2 \geq \int_{-\infty}^{u-k\sigma} (u-k\sigma-u)^2 \; f(x)dx + \int_{u+k\sigma}^{+\infty} (u+k\sigma-u)^2 \; f(x)dx$$

$$\geq \int_{-\infty}^{u-k\sigma} k^2\sigma^2 \; f(x)\,dx + \int_{u+k\sigma}^{+\infty} k^2\sigma^2 \; f(x)\,dx$$

$$\geq k^2\sigma^2 \left[\int_{-\infty}^{u-k\sigma} f(x)\,dx + \int_{u+k\sigma}^{+\infty} f(x)\,dx\right] .$$

The first integral will be recognized as the probability that x will be smaller than u-kσ, and the second integral, the probability that x will exceed u+kσ. The inequality can therefore be written

$$\sigma^2 \geq k^2\sigma^2 \left[P_r\,(x < u-k\sigma) + P_r\,(x > u+k\sigma)\right]$$

$$\geq k^2\sigma^2 \left[P_r\,(x-u < -k\sigma) + P_r\,(x-u > k\sigma)\right]$$

$$\geq k^2\sigma^2 \left[P_r\,(|x-u| > k\sigma)\right] ,$$

or finally, $P_r\,(\,|\,x-u\,|\, > k\sigma) \leq 1/k^2$.

This is Tchebycheff's inequality, which simply states that the probability is equal to or less than $1/k^2$ that a randomly drawn sample observation will lie farther than k population standard deviations from the population mean. It can be applied to an entire sample of n observations by substituting the sample mean for x and the true standard error of the mean for σ. The statement then becomes: the probability does not exceed $1/k^2$ that the mean of a random sample will lie farther than k standard errors of the mean from the population mean.

The inequality uses both the mean and variance of the population. If either is known, it can be substituted into the inequality along with an hypothesized value for the other parameter and the observed value x. The inequality can then be used to test the hypothesis which determined the value substituted for the unknown parameter. The significance level, $\propto$, must equal $1/k^2$, so $k = \frac{1}{\sqrt{\propto}}$ and the hypothesis is rejected at the $\propto$ level of significance

if $\frac{|x-u|}{\sigma} > \frac{1}{\sqrt{\propto}}$. Again, $\bar{x}$ and $\sigma_{\bar{x}}$ can be substituted for x and

σ to make the test applicable to samples of more than one observation. Obviously, the inequality can also be used for prediction and for the setting of tolerance limits.

Tchebycheff's inequality suffers from a number of deficiencies. First it is distribution-free only in the limited sense that it does not completely specify the shape or contour of the distribution to which it is to apply. It does, however, require a knowledge of the population variance, (or true standard error of the mean), which is seldom available in the absence of knowledge of the distribution's form. Second, it is, by nature, inexact; the only explicit probability statement that can be made concerns the upper bound for a probability rather than the probability itself. Finally, it is a weak test in that, when applied to small samples, it is generally unlikely to reject a false null hypothesis unless the hypothesis is spectacularly in error; small discrepancies between null hypothesis and true condition are usually detected only when extremely large samples are taken. This last shortcoming could have been predicted on the basis of the derivation. The central term

$\int_{u-k\sigma}^{u+k\sigma} (x-u)^2 f(x)\,dx$, was completely discarded and the values

of x which would minimize the two remaining integrals were substituted into them. The net result is that the term σ^2 on the left of the inequality sign is, in all probability, much greater than the sum of the terms constituting the right hand side of the inequality. The weakness of the test could also be predicted on the basis of the fact that an otherwise unknown distribution is poorly described by a mere knowledge of its variance or mean or any other single parameter.

Despite these weaknesses, the inequality has as much strength as can be obtained under the assumed conditions. That is to say, the discrepancy between the values on opposite sides of the inequality sign cannot be reduced without the imposition of further restrictions(6). Many "stronger" Tchebycheff-like inequalities have been developed at the cost of introducing more and more elaborate assumptions about the population distribution (such as requiring that the distribution be unimodal or symmetrical or that it increase monotonically in progressing from its tails to its mode, etc.). This means, of course, that the proper use of such inequalities is restricted to populations about whose distributions more and more is known. It seems to be in the nature of inequalities, therefore, that strength and freedom from assumptions are inversely related.

Despite its weakness as a statistical test, Tchebycheff's inequality has played an important part in the mathematical development of probability theory. It has been extended to bivariate (1, 3, 7) and multivariate (4, 7) distributions and has been mathematically "generalized" so as to include a wide variety of inequalities as special cases. It is involved in many important statistical derivations. However, it is rarely used now as a statistical test. Pearson (18) sums up what is probably still the prevailing attitude toward Tchebycheff inequalities as statistical tests: "On the whole we must express disappointment at the results of Tchebycheff's process. We had found Tchebycheff's own limit based on the second moment of small practical value, although it is to be found occupying a prominent position in many continental works on probability. By extending it to higher moments and product-moments we have reached results which are great improvements on the original Tchebycheff limit, but the method still lacks the degree of approximation (except for probabilities over .99, say) which would make the results of real value in practical statistics."

BIBLIOGRAPHY

1. Berge, P. O., A note of a form of Tchebycheff's theorem for two variables. Biometrika, 1937, 29, 405-406.

2. Birnbaum, Z. W., On random variables with comparable peakedness. Annals of Mathematical Statistics, 1948, 19, 76-81.

3. Birnbaum, Z. W., Raymond, J. and Zukerman, H. S., A generalization of Tchebychev's inequality to two dimensions. Annals of Mathematical Statistics, 1947, 18, 70-79.

4. Camp, B. H., Generalization to N dimensions of inequalities of the Tchebycheff type. Annals of Mathematical Statistics, 1948, 19, 568-574.

5. Craig, C. C., On the Tchebycheff inequality of Bernstein. Annals of Mathematical Statistics, 1933, 4, 94-102.

6. Cramér, H., Mathematical methods of statistics, Princeton, N. J.: Princeton University Press, 1946, pp. 182-183.

7. GODWIN, H. J., On generalizations of Tchebychef's inequality. Journal of the American Statistical Association, 1955, 50, 923-945.

8. Guttman, L., A distribution-free confidence interval for the mean. Annals of Mathematical Statistics, 1948, 19, 410-413.

9. Guttman, L., An inequality for kurtosis. Annals of Mathematical Statistics, 1948, 19, 277-278.

10. HOEL, P. G., Introduction to mathematical statistics, New York: Wiley, 1947, 172-173.

11. Hsu, P. L., The approximate distributions of the mean and variance of a sample of independent variables. Annals of Mathematical Statistics, 1945, 16, 1-29.

12. Kolmogorov, A., Foundations of the theory of probability, New York: Chelsea, 1950, 42-43.

13. Leser, C. E. V., Inequalities for multivariate frequency distributions. Biometrika, 1942, 32, 284-293.

14. Lurquin, C., Sur le criterium de Tchebycheff. Comptes Rendus (Paris), 1922, 175, 681-683.

15. Meidell, B., Sur la probabilité des erreurs. Comptes Rendus (Paris), 1923, 176, 280-282.

16. von Mises, R., The limits of a distribution function if two expected values are given. Annals of Mathematical Statistics, 1939, 10, 99-104.

17. Narumi, S., On further inequalities with possible application to problems in the theory of probability. Biometrika, 1923, 15, 245-253.

18. Pearson, K., On generalized Tchebycheff theorems in the mathematical theory of statistics. Biometrika, 191 9, 12, 284-296.

19. Smith, C. D., On generalized Tchebycheff inequalities in mathematical statistics. American Journal of Mathematics, 1930, 52, 109-126.

20. SMITH, C. D., Tchebycheff inequalities as a basis for statistical tests. Mathematics Magazine, 1955, 28, 185-195.

21. Smith, C. D., On Tchebycheff approximation for decreasing function. Annals of Mathematical Statistics, 1939, 10, 190-192.

22. Shohat, J. A., Inequalities for moments of frequency functions and for various constants. Biometrika, 1929, 21, 361-375.

23. Wald, A., Generalization of the inequality of Markoff. Annals of Mathematical Statistics, 1938, 9, 244-255.

24. Winsten, C. B., Inequalities in terms of mean range. Biometrika, 1946, 33, 283-295.

CHAPTER XV

EXTREME VALUE DISTRIBUTIONS

The distribution of the largest, or smallest, value in a sample of n observations has been investigated by Fisher and Tippett (4), Gumbel (5-9, 15) and others (2, 13, 14). These investigations have met with qualified success: the distribution of an extreme sample value has been obtained for samples of infinite size from certain types or classes of population. Gumbel has investigated and tabled (9, 15) extreme values and near-extreme values for large samples from populations whose distribution is of the exponential type, "which covers, among others, the exponential, the normal, and the chi-square distribution." Extreme value distributions find important application in predicting the "return period" for floods and other meteorological phenomena, and in strength of materials investigations since it is the weakest of n "fibers", the worst of n flaws, or the heaviest of n loads which determine when and where fracture will begin.

If a very large sample is taken, the correlation between the largest and smallest sample values becomes negligible (5) and the sample extremes may be regarded as effectively independent. Under these circumstances the distribution of the sample range can be obtained from the joint distribution of the two extremes. The distribution of the range, obtained in this way, necessarily incorporates all assumptions made in obtaining the distributions for the extremes. Gumbel (6) has tabled probabilities for ranges and range-like statistics for samples from an "unlimited symmetrical initial distribution of the exponential type."

It is clear that the extreme value statistics discussed above and range statistics derived from them are completely valid only for infinitely large samples. Furthermore they are distribution-free only in the very limited sense that the form of the underlying population need not be known fully but rather need be known only to the degree necessary positively to categorize it as belonging to a certain specified class of populations. Such restrictions place these statistics outside the scope of this report and no attempt will be made to describe them in detail.

BIBLIOGRAPHY

1. Cramer, H., Mathematical methods of statistics, Princeton: Princeton University Press, 1946, 367-372.

2. Dodd, E. L., The greatest and least variate under general laws of error. Transactions of the American Mathematical Society, 1923, 25, 525-539.

3. Epstein, B., Application of the theory of extreme values in fracture problems. Journal of the American Statistical Association, 1948, 43, 403-412.

4. FISHER, R. A. and TIPPETT, L. H. C., Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 1928, 24, 180-190.

5. Gumbel, E. J., On the independence of the extremes in a sample. Annals of Mathematical Statistics, 1946, 17, 78-81.

T 6. Gumbel, E. J., Probability tables for range. Biometrika, 1949, 36, 142-148.

7. Gumbel, E. J., Ranges and midranges. Annals of Mathematical Statistics, 1944, 15, 414-422.

8. Gumbel, E. J., The distribution of the range. Annals of Mathematical Statistics, 1947, 18, 384-412.

T 9. Gumbel, E. J. and Greenwood, J. A., Table of the asymptotic distribution of the second extreme. Annals of Mathematical Statistics, 1951, 22, 121-124.

10. Kendall, M. G., Note on the distribution of quantiles for large samples. Journal of the Royal Statistical Society (B), 1940, 7, 83-85.

11. Kimball, B. F., On the problem of forecasting extreme values from a curve fitted to the type I extreme-value distribution, (Mimeographed)Address delivered at New York meeting if I.M.S, Dec. 27, 1955.

12. Kimball, B. F., Practical applications of the theory of extreme values. Journal of the American Statistical Association, 1955, 50, 517-528.

13. Lieblein, J., A new method of analysing extreme value data. Technical Note No. 3053, National Advisory Committee Aeronautics, 1954.

14. Lieblein, J., On the exact evaluation of the variances and covariances of order statistics in samples from the extreme-value distribution. Annals of Mathematical Statistics, 1953, 24, 282-287.

T 15. National Bureau of Standards, Probability tables for the analysis of extreme-value data. National Bureau of Standards, Applied Mathematics Series No. 22, Washington, D. C.: U. S. Government Printing Office, 1953.

16. WILKS, S. S., Order statistics. Bulletin of the American Mathematical Society, 1948, 54, 6-50.

CHAPTER XVI

OBTAINING AN OVERALL PROBABILITY FOR SEVERAL INDEPENDENT TESTS

It is sometimes desirable to obtain an overall probability for a number of separate and independent tests of the same null hypothesis. It may be that conducting an additional single test upon the aggregate data is justifiable theoretically but not practically because it would require excessive labor or delay. Or it may be that the data cannot properly be combined. For example, one test may have been a t-test for matched pairs, another a t-test without matching, a third, the sign test, etc.

What is desired is the probability of acquiring by chance a set of test outcomes as extreme as, or more extreme than, those actually obtained. This overall probability is not the product of the probabilities of the individual tests. To illustrate, if each of five tests yields results at the .50 level, the product of the five probabilities is .03125, although it is clear that in combination the five tests are even less suggestive of a false null hypothesis than they are individually. Probabilities can range from zero to one. Each time a probability is added to a set of probabilities, the product of the probabilities must diminish (or remain the same if the added probability is 1).

The not uncommon, but fallacious, belief that the overall probability for a set of test outcomes is expressed by the product of the individual probabilities is apparently due to misinterpretation of compound probability. If events A, B, C, yield the completely independent outcomes a, b, c, whose individual chance probabilities are p_a, p_b, p_c, then the product $p_a p_b p_c$ gives the a priori probability that outcome a will result for event A, outcome b will result for event B, and outcome c will result for event C. Such a procedure is invalid for the combination of test probabilities for two reasons. First, we are not interested in the probability that test A will yield probability p_a, test B will yield probability p_b, and test C will yield probability p_c. Rather we are interested in the probability that the probabilities p_a, p_b,

p_c will be obtained, each probability applying to some unspecified one of the three tests. Second, test probabilities are cumulative probabilities and therefore do not express the probability of a single, obtained, outcome, but rather the probability of the obtained outcome plus the probabilities of all of a defined class of less "expected", and unobtained, outcomes. Since all of the outcomes referred to by the smallest test probability are also, in a sense, referred to by a portion of each of the other test probabilities, the requirement of independence has not been met. Multiplying test probabilities therefore not only does not give us the probability we seek; it is not even a valid procedure for obtaining a probability which we do not seek.

There are several methods of obtaining overall probabilities. Each requires that the component tests be independent and test the same null hypothesis. The requirement of independence means that if the null hypothesis is true there is no common underlying factor in any of the data upon which the individual tests are based which would tend to produce similar test outcomes. Specifically this means that unless the tests are statistically independent (which is usually not the case) they must have been conducted upon separate and nonoverlapping sets of data yielded by separate and nonoverlapping groups of subjects (unless the null hypothesis is confined to the population of tested subjects). The reason for the requirement that the individual tests must test the same null hypothesis is obvious.

The rationale of the binomial, or Wilkinson, method is as follows. If, before collecting data, it is decided to use the same significance level, $\propto$, for each of N independent tests, then each test must have one of two outcomes: significance or insignificance. Significance is therefore binomially distributed, with probability $\propto$ on a single trial. The probability that n or more of N independent tests will yield probabilities falling within the significance level $\propto$ is then $\sum_{r=n}^{N} \binom{N}{r} \propto^{r} (1-\propto)^{N-r}$. This probability can be obtained from tables of cumulative binomial probabilities or from tables (19) or graphs (17) designed expressly for this purpose. A less desirable solution is afforded by the normal approximation to the binomial. This is justified only in those rare cases for which the binomial tables are not sufficiently extensive. The normal tables are entered with the critical ratio $\frac{|n - N\alpha| - .5}{\sqrt{N\alpha\,(1-\alpha)}}$, a one-tailed test being conducted, with small probabilities corresponding to values of n greater than $N\alpha$. The approximation cannot be expected to be good if N is small (say less than 20) or if $N\alpha$ is less than 5.

The binomial method presupposes that the size of the rejection region, i.e. the significance level, for each test was selected prior to collection of data, and that the same level α was selected for each test. The selection of α prior to collection of data insures an absence of a posteriori bias in obtaining an overall significance level. The binomial method also requires that each of the N tests be capable of of an outcome whose cumulative probability is exactly α. That is to say, the test statistic need not be continuously distributed, but, if not, it must have a discrete value corresponding to exactly the α level of significance not simply falling within the level α. Otherwise the binomial method would be inaccurate in the direction of conservatism: it would fail to announce significance as frequently as it occurred.

If the experimenter knows that nonchance values of the test statistic can only fall on one tail of its distribution, or if he is only interested in nonchance results falling on a specified tail, he will use the one-tailed α level of significance for all N tests and the binomial method will be highly appropriate for their combination. However, if so far as the experimenter knows, nonchance results can fall on either one of the two tails, and if he is interested in both eventualities, the binomial method becomes ambiguous in interpretation.

Ordinarily one uses a two-tailed test when one is unable to predict the direction of nonchance results. When a single test is conducted this is a reasonable and uncomplicated procedure. However, even though the experimenter may be unable to specify on which "side" of a false null hypothesis the true condition will lie, it cannot lie on both sides, and the result would therefore be highly ambiguous if probabilities for two-tailed tests were "combined" by the binomial method. If he uses the two-tailed α for each of his N tests, the binomial method still tells him precisely the probability that n of his tests will yield probabilities within the two-tailed α region by chance. However, the usual supposition that if the chance probability is small the tested effects must be due to some nonchance factor may become, in this case, a non sequitur. For example, the chance probability that of 28 tests, 4 or more will yield results significant at the two-tailed .05 level is .049, and an experimenter using the .05 level for his overall significance level would reject the null hypothesis. However, what is the "true" hypothesis if two of the four significant tests fell in the $\frac{1}{2}\alpha$ region and the other two fell on the opposite tail in the $1 - \frac{1}{2}\alpha$ region? The experimenter obviously cannot properly select a one-tailed significance region in advance of all data collection if he does not know in what direction to expect departures from the hypothesized condition. On the other hand, if he selects a one-tailed region on the basis of examination of the data, he is guilty of introducing an a posteriori bias, and the alleged overall probability for his results will not be the true probability. If he insists upon combining two-tailed tests, he will be able to make a precise probability statement about a chance event, which probability has little bearing on the nonchance event in which he is really interested.

The Chi-Square, or Fisher, method gives the probability of obtaining a certain product for the one-tailed cumulative probabilities of several tests. While the overall probability of a series of tests is not expressed by the product of their separate probabilities, that product has, itself, a probability of occurrence which can be regarded as the overall probability for the series of tests.

Expanding a treatment and derivation given more concisely by Wallis (18), let N be the number of tests whose probabilities, p_1, p_2 ..., p_N, are to be combined. Let each test be capable of yielding any cumulative probability between 0 and 1, each value being equally likely, i.e., assume each test statistic to be continuously distributed. Then the sample space for the product $p_1\,p_2\,....\,p_N = k$ is a square when N = 2, a cube when N = 3, and an equal sided, N-dimensional solid when N > 3. When N = 2, the probability that the product $p_1\,p_2$ does not exceed some value k is that area in a square of unit edge which lies on the convex side of the equilateral hyperbola $p_1 p_2 = k$ (See Figure 5). It is therefore one minus the area enclosed by the hyperbola. The enclosed area is $\int_k^1 (1-p_1)\,dp_2$. Substituting

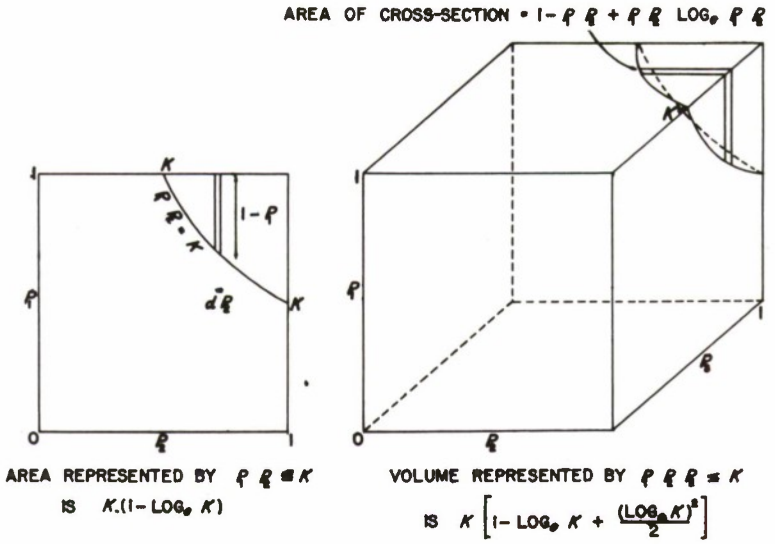

AREA OF CROSS-SECTION = $1 - p\,q + p\,q\ \mathrm{LOG}_e\ p\,q$
K
1
$p\,q = K$
$1 - p$
p
dq
K
0
q
1
AREA REPRESENTED BY $p\,q \leqq K$
IS $K.(1 - \mathrm{LOG}_e\ K)$
K
1
p
1
q
0
q
VOLUME REPRESENTED BY $p\,q\,q \leqq K$
IS $K\left[1 - \mathrm{LOG}_e\ K + \frac{(\mathrm{LOG}_e\ K)^2}{2}\right]$

Figure 5

k/p_2 for p_1, this becomes $\int_k^1 (dp_2 - k\,\frac{dp_2}{p_2})$ or $1 - k + k \ln k$, and the desired probability is 1 minus this or $k(1 - \ln k)$. When N = 3, the desired probability is the volume of a unit cube minus that portion whose cross section is $1 - p_1 p_2 + p_1 p_2 \ln p_1 p_2$, (see above) and whose perpendicular dimension extends from $p_3 = k$ to $p_3 = 1$.

The volume to be subtracted is $\int_k^1 (1 - p_1 p_2 + p_1 p_2 \ln p_1 p_2)\,dp_3$ or, substituting k/p_3 for $p_1 p_2$,

$$\int_k^1 \left(1 - \frac{k}{p_3} + \frac{k}{p_3} \ln \frac{k}{p_3}\right) dp_3 = 1 - k + k \ln k + \int_k^1 k\,\frac{dp_3}{p_3} \ln \frac{k}{p_3}$$

the remaining integral, integrated by parts, becomes ${}_k^1\left[k \ln \frac{k}{p_3} \ln p_3\right] - k \int_k^1 (\ln p_3)\left(\frac{-k\,p_3^{-2}}{k p_3^{-1}}\right) dp_3$

$= k \int_k^1 \frac{dp_3}{p_3} \ln p_3 = k\;{}_k^1\left[\frac{(\ln p_3)^2}{2}\right] = \frac{-k\,(\ln k)^2}{2}$. The subtracted volume, then, is $1 - k + k \ln k - \frac{k(\ln k)^2}{2}$. And the desired probability is $k\left[1 - \ln k + \frac{(\ln k)^2}{2}\right]$. The general term for the probability of the product N independent tests is, then

$$k \sum_{r=0}^{N-1} \frac{(-\ln k)^r}{r\,!} \quad \text{which can be written as} \quad \sum_{r=0}^{N-1} \frac{e^{\ln k}(-\ln k)^r}{r!}$$

which is the sum of the first N terms of the Poisson distribution whose mean is $-\ln k$. However, it is known that the probability for a value of χ^2 based on 2N degrees of freedom is given by the sum of the first N terms of the Poisson whose mean is $\chi^2/2$. Therefore the probability $\sum_{r=0}^{N-1} \frac{e^{\ln k}(-\ln k)^r}{r!}$ of the product k is also the probability of that value of χ^2 based on 2N degrees of freedom for which $\chi^2/2 = -\ln k$. Stated differently, when based on 2N degrees of freedom $\chi^2 = -2 \ln k$ has the probability we seek.

Therefore to obtain the overall probability for N tests whose separate probabilities yield the produce k, enter the χ^2 tables with 2N degrees of freedom and find the probability for the value of χ^2 equal to $-2 \ln k$. This is the probability of the product k.

An alternative and equivalent method does not require the evaluation of logarithms. As mentioned earlier, the probability for for a value of χ^2 with 2N degrees of freedom is the sum $\sum_{r=0}^{N-1} \frac{e^{-\frac{\chi^2}{2}} (\frac{\chi^2}{2})^r}{r!}$. For two degrees of freedom, N=1 and the probability becomes simply $e^{-\frac{\chi^2}{2}}$. Solving $p = e^{-\frac{\chi^2}{2}}$ for χ^2, we have $-\frac{\chi^2}{2} = \ln p$ or $\chi^2 = -2 \ln p$. That is to say, the value of any χ^2 based on two degrees of freedom is minus twice the natural logarithm of its own probability. Phrased differently, one can obtain minus twice the natural logarithm of any probability by entering the chi-square tables with that probability and with two degrees of freedom and reading off the corresponding value of chi square. Suppose this is done for each of the N probabilities for which the overall probability is sought. Then for each probability p_i, we obtain a χ_i^2 for 2 d.f. $= -2 \ln p_i$. Because of the additive property of χ^2, these values of χ^2 based on two degrees of freedom can be summed to give a total value of χ^2 based on the sum of the separate degrees of freedom.

$$\chi_i^2 \text{ for 2 d.f.} = -2 \ln p_i$$

$$\sum_{i=1}^{N} (\chi_i^2 \text{ for 2 d.f.}) = \sum_{i=1}^{N} -2 \ln p_i = -2 \sum_{i=1}^{N} \ln p_i$$

$$\text{total } \chi^2 \text{ based on 2N d.f.} = -2(\ln p_1 + \ln p_2 + .. \ln p_N)$$

$$= -2 \ln(p_1 p_2 \ldots p_N) = -2 \ln k.$$

Therefore, the total χ^2 based on 2N degrees of freedom has precisely the probability we seek, and this total χ^2 has been obtained without resort to any tables of logarithms. Extensive tables of χ^2 for two degrees of freedom (8) have been provided for use with this method. Graphs (1, 2) exist which give the probability of the product of two probabilities. In using the chi square method, Yates' correction should never be applied as it is completely inappropriate. Also, each of the individual probabilities to be combined must be continuously distributed, i.e., the "population" probability must be capable of assuming any value between zero and one. This, in turn, means that the test statistic must be continuously distributed, which eliminates many distribution-free tests. If the test statistic is capable of assuming a large number of different values, however, the technique may be used as an approximate method. Another requirement of the chi-square method is that the probabilities to be combined must be exact cumulative probabilities, not simply "significance levels" within which the cumulative probability has fallen. Thus the experimenter must have available tables of exact cumulative probabilities for each of the test statistics whose probabilities are to be combined; tables giving the values of the test statistic at the conventional significance levels such as .10, .05, .01, .001 will not suffice unless linear interpolation is performed and unless it yields very nearly exact values. A further requirement is that the cumulative probabilities used for the individual tests must all be one-tailed probabilities, with those probabilities near zero all implying the same type of departure from the hypothesized condition and with those probabilities near one all implying the opposite type of departure. If the experimenter wishes to conduct a two-tailed <u>overall</u> test at the significance level α, he simply rejects the null hypothesis if the product k is either so small that its probability is less than $\frac{1}{2}\alpha$ or so large that its probability is greater than $1 - \frac{1}{2}\alpha$. Thus the chi-square method is free of the ambiguity surrounding the binomial method when a two-tailed overall test is required.

Wallis (18) has outlined the method of obtaining the probability of a product of individual probabilities when some of them are discretely distributed.

There are other methods of obtaining overall probabilities. A technique somewhat similar to that described as the binomial method is attributed (3, p. 562) to Tippett. A technique (5, 6, 12, 13, 14, 15, 16) which is essentially the chi-square method was discovered subsequently but independently by Karl Pearson. Birnbaum (3) states that "no single method of combining independent tests of significance is optimal in general, and hence ... the kinds of tests to be combined should be considered in selecting a method of combination." Various methods are examined by him in (3), the two methods described above, i.e., the binomial method and the chi-square method apparently being most effective in the generality of applications.

SUMMARY

Two methods have been described for obtaining an overall probability for the outcomes of a set of statistical tests, using as "data" the obtained cumulative probabilities for each of the individual outcomes. Both methods require that the component tests be independent and test the same null hypothesis.

The binomial method gives the probability that of N tests, n or more will yield cumulative probabilities falling within a predesignated significance level α. The individual test statistics neednot be continuously distributed; however, each must have a value corresponding to a cumulative probability of exactly α. The binomial method is highly appropriate when the individual tests to be combined are one-tailed, and a one-tailed overall test of the null hypothesis is required. If α is taken as a two-tailed significance level, the binomial method remains mathematically valid, giving the chance probability of the obtained results. However, small chance probabilities can no longer be taken as presumptive evidence that the null hypothesis is false, since they do not necessarily imply the existence of a more likely alternative. If nearly equal proportions of the n significant tests fall on opposite tails $\frac{1}{2}\alpha$ and $1 - \frac{1}{2}\alpha$, then rejection of the null hypothesis is unjustified since no alternative hypothesis accounts for the results any better than does the null hypothesis.

While the overall probability of a series of tests is not expressed by the product of their separate probabilities, that product has, itself, a probability of occurrence which can be regarded as the overall probability for the series of tests. The chi-square method gives the cumulative probability for the product of the one-tailed cumulative probabilities of N tests. It requires: (a) that the individual test statistics be continuously distributed, i.e., that every cumulative probability from zero to one be equally likely, (b) that one-tailed cumulative probabilities be used for the individual tests, and that a cumulative probability on a given side of .50 imply the same direction of deviation from H_o for every test, (c) that for each test the exact cumulative probability be used, not simply the "significance level" within which that cumulative probability fell. The overall test of significance may be made two-tailed at the α level of significance by rejecting the null hypothesis if the one-tailed cumulative probability of the product falls either between zero and $\frac{1}{2}\alpha$ or between $1 - \frac{1}{2}\alpha$ and 1.

For specific cases, the following table may be helpful in deciding which of the two methods is appropriate.

Restrictive Conditions	Method	
	Binomial	Chi-Square
Continuously distributed test statistics required?	No*	Yes
Exact cumulative probabilities required?	No*	Yes
Two-tailed tests are ambiguous?	Yes	No

*See text for qualifications.

BIBLIOGRAPHY

T 1. Baker, P. C., Combining tests of significance in cross-validation. Educational and Psychological Measurement , 1952, 12, 300-306.

T 2. Baker, P. C., The probability of two combined tests of significance, Graph published by Purdue University.

3. Birnbaum, A., Combining independent tests of significance. Journal of the American Statistical Association, 1954, 49, 559-574.

4. Brozek, J. and Tiede, K., Reliable and questionable significance in a series of statistical tests. Psychological Bulletin, 1952, 49, 339-341.

T 5. David, Florence N., On the Pλ_n test for randomness: remarks, further illustration, and table of Pλ_n for given values of -log$_{10}$ λ_n. Biometrika, 1934, 26, 1-11.

6. David, Florence N. and Johnson, N. L., The probability integral transformation when parameters are estimated from the sample. Biometrika, 1948, 35, 182-190.

7. Fisher, R. A., Statistical methods for research workers, New York: Hafner, 1954, pp. 99-101.

T 8. Gordon, M. H., Loveland, E. H. and Cureton, E. E., An extended table of chi-square for two degrees of freedom, for use in combining probabilities from independent samples. Psychometrika, 1952, 17, 311-316.

9. Jones, L. V. and Fiske, D. W., Models for testing the significance of combined results. Psychological Bulletin, 1953, 50, 375-382.

10. Lancaster, H. O., The combination of probabilities arising from data in discrete distributions. Biometrika, 1949, 36, 370-382.

11. Mood, A. M., Introduction to the theory of statistics, New York: McGraw-Hill, 1950, pp. 107-108.

12. Nair, K. R., A note on the exact distribution of λ_n. Sankhyā, 1937, 3, 171-174.

13. Pearson, E. S., On questions raised by the combination of tests based on discontinuous distributions. Biometrika, 1950, 37, 383-398.

14. Pearson, E. S., The probability integral transformation for testing goodness of fit and combining independent tests of significance. Biometrika, 1938, 30, 134-148.

15. Pearson, K., On a new method of determining 'goodness of fit'. Biometrika, 1934, 26, 425-442.

* 16. Pearson, K., On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 1933, 25, 379-410.

T 17. Sakoda, J. M., Cohen, B. H. and Beall, G., Test of significance for a series of statistical tests. Psychological Bulletin, 1954, 51, 172-175.

18. WALLIS, W. A., Compounding probabilities from independent significance tests. Econometrica, 1942, 10, 229-248.

T 19. Wilkinson, B., A statistical consideration in psychological research. Psychological Bulletin, 1951, 48, 156-158.

UNCLASSIFIED

Wright Air Development Division, Aerospace
Medical Division, Wright-Patterson Air
Force Base, Ohio.
DISTRIBUTION-FREE STATISTICAL TESTS,
by James V. Bradley. August 1960. 386p.
incl. illus., refs. (Proj. 7184; Task 71581)
WADD TR 60-661 Unclassified report

As a result of an extensive survey of the literature, a large number of distribution-free statistical tests are examined. Tests are grouped together primarily according to general type of mathematical derivation or type of statistical information used in conducting the test. Each of the more important tests

(over)

is treated under the headings: Rationale, Null Hypothesis, Assumptions, Treatment of Ties, Efficiency, Application, Discussion, Tables, and Sources. Derivations are given and mathematical interrelationships among the tests are indicated. Strengths and weaknesses of individual tests, and of distribution-free tests as a class compared to parametric tests are discussed.

UNCLASSIFIED

UNCLASSIFIED

UNCLASSIFIED

UNCLASSIFIED

Wright Air Development Division, Aerospace
Medical Division, Wright-Patterson Air
Force Base, Ohio.
DISTRIBUTION-FREE STATISTICAL TESTS,
by James V. Bradley. August 1960. 386p.
incl. illus., refs. (Proj. 7184; Task 71581)
WADD TR 60-661 Unclassified report

As a result of an extensive survey of the literature, a large number of distribution-free statistical tests are examined. Tests are grouped together primarily according to general type of mathematical derivation or type of statistical information used in conducting the test. Each of the more important tests

(over)

is treated under the headings: Rationale, Null Hypothesis, Assumptions, Treatment of Ties, Efficiency, Application, Discussion, Tables, and Sources. Derivations are given and mathematical interrelationships among the tests are indicated. Strengths and weaknesses of individual tests, and of distribution-free tests as a class compared to parametric tests are discussed.

UNCLASSIFIED

UNCLASSIFIED

UNCLASSIFIED

Wright Air Development Division, Aerospace
Medical Division, Wright-Patterson Air
Force Base, Ohio.
DISTRIBUTION-FREE STATISTICAL TESTS,
by James V. Bradley. August 1960. 386p.
incl. illus., refs. (Proj. 7184; Task 71581)
WADD TR 60-661 Unclassified report

As a result of an extensive survey of the literature, a large number of distribution-free statistical tests are examined. Tests are grouped together primarily according to general type of mathematical derivation or type of statistical information used in conducting the test. Each of the more important tests

(over)

UNCLASSIFIED

UNCLASSIFIED

is treated under the headings: Rationale, Null Hypothesis, Assumptions, Treatment of Ties, Efficiency, Application, Discussion, Tables, and Sources. Derivations are given and mathematical interrelationships among the tests are indicated. Strengths and weaknesses of individual tests, and of distribution-free tests as a class compared to parametric tests are discussed.

UNCLASSIFIED

UNCLASSIFIED

Wright Air Development Division, Aerospace
Medical Division, Wright-Patterson Air
Force Base, Ohio.
DISTRIBUTION-FREE STATISTICAL TESTS,
by James V. Bradley. August 1960. 386p.
incl. illus., refs. (Proj. 7184; Task 71581)
WADD TR 60-661 Unclassified report

As a result of an extensive survey of the literature, a large number of distribution-free statistical tests are examined. Tests are grouped together primarily according to general type of mathematical derivation or type of statistical information used in conducting the test. Each of the more important tests

(over)

UNCLASSIFIED

UNCLASSIFIED

is treated under the headings: Rationale, Null Hypothesis, Assumptions, Treatment of Ties, Efficiency, Application, Discussion, Tables, and Sources. Derivations are given and mathematical interrelationships among the tests are indicated. Strengths and weaknesses of individual tests, and of distribution-free tests as a class compared to parametric tests are discussed.

UNCLASSIFIED

UNCLASSIFIED